Design, Ligand Binding and Folding of Ankyrin and Armadillo Repeat Proteins Studied by Solution NMR Spectroscopy

---

Dissertation
zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der
Universität Zürich

von
Christina Ewald

aus
Deutschland

Promotionskomitee
Prof. Dr. Oliver Zerbe (Leitung der Dissertation)
Prof. Dr. Andreas Plückthun
Prof. Dr. Stephan Grzesiek
Prof. Dr. Frédéric Allain

Zürich, 2013

Elementary, My Dear Watson.

Erklärung

Diese Dissertation wurde selbständig, ohne unerlaubte Hilfe im Sinne der Promotionsordnung vom 03. Februar 2005 angefertigt. Bei der Abfassung der Dissertation wurden keine anderen als die darin angegebenen Hilfsmittel benutzt.

Zürich, August 2013

Christina Ewald

Acknowledgements

The completion of a PhD thesis is rarely a singular effort, and this work has been no exception to this rule. It is impossible to spend five years working on a collaborative project without interacting and working with a large number of people. A number of individuals, groups, and organisations have contributed to the success of this work and I would like to thank these people for their valuable contributions.

Firstly, I would like to thank Prof. Dr. Oliver Zerbe and Prof. Dr. Andreas Plückthun for giving me this opportunity and their continued support and guidance throughout this thesis.
I am also greatful to the other members of my thesis committee Prof. Dr. Stephan Grzesiek and Prof. Dr. Frédéric Allain for their time and helpful suggestions.

I am thankful for funding and financial support from the SNF (Sinergia Project Grant No. CRSI00_122686) and the Hartmann-Müller-Hahn Foundation.

To the members of the Zerbe, Plückthun, Robinson and Caflisch groups who have contributed scientific input, guidance and help, I am greatly thankful, especially for your openness and unreserved generosity with respect to the sharing of equipment and expertise in times of need.

Specifically I would like to thank Randall Watson and Martin Poms for their tireless dedication to produce the best coffee on the Irchel campus, for their friendship, fruitful discussions and support in the "downstairs" office and Jacopo Marino for his balance, charisma and music.

Christian Reichen for his constant scientific support, enthusiasm and motivation unheard of before in the history of science, Reto Walser for his always patient explanations of NMR experiments, Fabian Bumbak for his questions and his very special self, the Italian fraction from the Caflisch and various other labs for interdisciplinary and entertaining lunch meetings and general party disorganisation, and Martin Christen for last minute custom fitting of excessive amounts of data.

I would also like to thank Simon Jurt and Nadja Bross for their continuous efforts to keep the arsenal of fickle technical equipment running and Salomé Fässler and Susanna Bachmann without whose guidance and efficient paperwork no student at the OCI and MLS Programme could have survived. Ralph Bollag from the mechanical workshop has contributed invaluably to the workings of our lab and his leaving is a great loss to the entire institute.

Finally, I want to thank my family for their constant long-distance support and trust to let me find my own way. Without you I would never have made it this far.

Let's see what's next…

Summary

Natural repeat proteins fulfil a plethora of important functions in cell biology like molecular recognition, cell adhesion and transport. Well known representatives of this protein class include armadillo, ankyrin, HEAT and tetratricopeptide repeat proteins. Proteins of this class constitute almost 20 % of proteins encoded in the human genome and contain tandem arrays of small, highly similar structural units. Several of these units stack against each other forming non-globular, elongated structures with long hydrophobic cores and extensive solvent exposed surfaces, determining topology and function of these proteins. Repeat proteins differ from globular proteins in several important characteristics. Firstly, they commonly display an extended solenoid fold. Secondly, they are mainly stabilized by short-range interactions between residues close in sequence, whereas the importance of long-range interactions for protein stability is greatly diminished compared to globular proteins.

Two repeat protein families were investigated during this PhD project – ankyrin repeat protein and armadillo repeat proteins. The natural ankyrin repeat is a very common type of motif and can be found in all three kingdoms spanning a wide range of functions, with the underlying theme being their ability to mediate protein-protein interactions by binding to three-dimensional epitopes. Armadillo repeat proteins are commonly involved in protein-protein or protein-peptide interactions, binding to peptides or unfolded parts of proteins. Importantly, their extended binding surface can bind peptides in extended conformation.

Protein engineering efforts aim at developing useful proteins with new or enhanced functions. The Plückthun group has undertaken an extensive design effort to create a highly stable designed consensus ankyrin repeat protein (DARPin) scaffold. These studies have cumulated in an optimized design, in which surface residues can be mutated to achieve binding to a desired target without compromising scaffold stability. Repeat proteins in general, and DARPins in particular, are an interesting subject upon which to study protein folding in order to understand the molecular base of their unusual stability. Their low contact order and modularity represents an intriguing background against which to study the mechanisms of protein folding and protein stability in a uniform environment.

In this work, we investigated the stability and folding behaviour of full-consensus designed ankyrin repeat proteins (DARPins) using a range of NMR, biophysical and computational experiments. The sequence background of identical repeats used for our study can be seen as a generalised example for the study of AR protein folding and enables the investigation of folding as a function of repeat number. Using proton-exchange methods in the presence and absence of chemical denaturation, we evaluated the stability of this ankyrin scaffold in a residue-resolved manner. In order to achieve this we had to first assign the backbone resonances of each repeat and the N- and C-terminal capping repeats – a problem which becomes progressively more difficult as additional internal repeat are added. Paramagnetic spin labels attached to either end of the proteins were successfully used to decrease ambiguitiy and allowed complete backbone resonance assignments.

Our results indicate that the stability of the full-consensus ankyrin repeat proteins is strongly dependent on the coupling between repeats. Some amide protons required more than a year to exchange at 37°C, highlighting the extraordinary stability of the proteins. Protein folding and stability data were analysed in the context of *in silico* predictions based on an Ising-type model. Denaturant induced unfolding, followed by deuterium exchange, chemical shift change and heteronuclear nuclear Overhauser effects, were consistent with an Ising-type description of equilibrium folding, in which the stability of each repeat to a first approximation depends on the number of neighbouring repeats and hence is largest for the central repeats. For native state deuterium exchange, we postulate local fluctuations to dominate exchange as unfolding events are too slow in these very stable proteins. The location of extraordinarily slowly exchanging protons indicates a very stable core structure in the DARPins which combines hydrophobic shielding with favorable electrostatic interactions. These investigations help the understanding of repeat protein architecture and the further design of DARPins for biomedical applications where high stability is required.

As pointed out above, armadillo repeat proteins are of particular interest due to their natural ability to bind peptides in extended conformation and thus recognize the amino acid sequence of the peptide rather than its tertiary structure. Using a repeat protein as a scaffold for generating engineered selective peptide binders can exploit their modular nature to produce specific peptide binders which could be customised simply by adding, removing or shuffling repeats in the same manner as the target peptide may be changed by increasing its length or changing its sequence. The key feature of this interaction, which forms the basis of our design efforts described in this thesis, is the conserved mode by which armadillo repeat proteins bind the backbone of peptides. Each repeat unit can bind two amino acid side chains through a conserved polar ladder - usually asparagines - on the armadillo domain. Interactions of the peptide side chains with the protein surface contribute further to the very high binding affinities observed in nature.

Initial proteins derived from a consensus design proved to be not sufficiently stable. In the search for a more stable consensus armadillo repeat protein scaffold for use in the directed evolution of peptide binders, and building on the work of Parmeggiani *et al.*, a molecular dynamics approach supported by NMR experiments and a variety of biophysical techniques has been employed to efficiently improve the thermodynamic stability of the N- and C-terminal capping repeats as well as the internal M-type repeat of this scaffold. Heteronuclear NMR revealed unfavorable interactions present at neutral but absent at high pH. MD-simulations suggested mutations in the capping repeats, which were then characterized by NMR, and by temperature- and denaturant-unfolding studies. The resulting proteins were monomeric species with cooperative folding and increased stability against heat and denaturant. This work indicates that *in silico* studies in combination with heteronuclear NMR and other biophysical tools may provide a basis for successfully selecting mutations that rapidly improve biophysical properties of the target proteins.

The assignment of chemical shifts of repeat proteins by NMR is very challenging due to the repetitive nature of their sequence. To facilitate this process we attempted segmental labelling using a split intein approach. Surprisingly, we observed that, when the repeat protein was expressed as two separate fragments with their intein ligation motifs present, the fragments showed an affinity for each other, although no peptide bond was formed. Removal of the split intein motifs resulted in the same observation, indicating that the

interaction was not mediated by the split intein. Thereby, we discovered the so-far unknown property of consensus designed armadillo repeat protein fragments to self-assemble. We feel that this knowledge may be very helpful to aid in the resonance assignment of complicated repeat proteins. As a proof of principle for this approach we performed structural, biophysical and thermodynamic studies characterizing this interaction using a full-consensus designed armadillo repeat protein with three internal repeats with the aim to structurally determine the self-assembly complex of the reconstituted protein, and to investigate the behaviour of uncomplexed armadillo repeat protein fragments. We analysed this interaction with reference to the structure of the entire protein, demonstrating that the same interfacial contacts are made. This feature of armadillo repeat proteins could be useful in the design of binding proteins from fragments, and a rather rapid engineering by exchanging non-covalent protein fragments.

The main project of this thesis was the characterisation of a previously selected peptide binder, VG_328, and the nature of its interaction with the neurotensin peptide. Again, spectral overlap due to repeating sequences in the selected 5-repeat protein VG_328 represented a severe technical challenge for detailed NMR analysis. Due to chemical shift degeneracy only a minor fraction of backbone resonances could be assigned by triple-resonance NMR. Additional annotations became possible when a series of truncated constructs were investigated. Although even in this case assignments were far from complete, chemical shift perturbations (CSPs) allowed recognition of the repeats involved in binding. Subsequent removal of repeats not involved in making contacts to the peptide resulted in a reduced-size binder with very similar affinity for NT, for which complete backbone assignments were achieved. Paramagnetic relaxation enhancement (PRE) studies using spin-labeled NT, and binding affinities of mutants provided experimental evidence which were compared with binding poses derived from MD simulations. All NMR data were compatible with the binding mode extracted from the MD trajectory. We postulate an antiparallel binding mode of the central part of NT to the upper part of the binding interface across helices 3 of all three internal repeats (peptide residues 7-11) involving numerous randomized protein positions. The N-terminal part of the peptide (residues 1-7) appears to sample a wide range of conformations, whereas the C-terminus of the peptide can form a stable salt bridge with a residue in helix 2 of the N-cap, thereby stabilizing the protein-peptide complex. This work was complemented by further mutational studies involving N-cap mutations and their effects on peptide binding and protein stability. These studies revealed the impact of a small set of mutations on the packing of the N-cap against the first internal repeat in detail, indicating a delicate balance between cap stabilization and peptide binding in this particular complex. This approach demonstrates that low-resolution NMR data were sufficient to guide further designs, and that the combination of NMR and MD data was successful in establishing the binding mode in the absence of crystallographic data even for a weakly binding ligand.

Especially, in early stages of such projects binding affinities are low, and protein binders may still contain flexible parts hampering crystallization. In this work we have developed a highly interdisciplinary approach combining mutagenesis, heteronuclear NMR spectroscopy and MD calculations as well as other biophysical tools. We believe that this approach can be a powerful strategy for analyzing difficult targets such as low affinity binders with multiple conformations and limited stability. We have also demonstrated that, even with information from limited NMR assignments, protein

sequences can be modified to yield proteins with superior characteristics that may eventually be amenable to high-resolution structural studies. Most importantly, this limited information is sufficient to drive the project forward and to verify original assumptions about the binding mode of the ligand in designed binders. In particular in the early stages of such projects it is of the utmost importance to ensure that the project is on the correct track.

Zusammenfassung

In der Natur übernehmen Repeat Proteine eine Fülle wichtiger zellbiologischer Aufgaben, zum Beispiel im Bereich der molekularen Erkennung, Zelladhäsion und des Proteintransports. Zu den bekannteren Mitgliedern dieser Proteinklasse zählen Armadillo, Ankyrin, HEAT und Tetratricopeptid Proteine. Nahezu 20 % der Proteine des menschlichen Genoms beihalten Proteine dieser Klasse, welche kleine, aneinandergereihte Einheiten von grosser struktureller Ähnlichkeit enthalten. Mehrere dieser Einheiten können nebeneinander gereiht längliche, nicht globuläre Proteine bilden. Ihr langestrecktes hydrophobes Zentrum und ihre grosse Oberfläche bestimmen die Gestalt und Eigenschaften dieser Proteine. Repeat Proteine unterscheiden sich von globulären Proteinen in mehreren wichtigen Eigenschaften. Sie besitzen üblicherweise eine langgestreckte Struktur, welche hauptsächlich durch Interaktionen zwischen Aminosäuren stabilisiert wird, die in der Proteinsequenz nahe beieinander liegen. Weitreichende räumliche Wechselwirkungen spielen für die Stabilität dieser Proteine nur eine untergeordnete Rolle im Vergleich zu globulären Proteinen.

Während dieser Doktorarbeit wurden zwei verschiedene Repeat Protein Familien untersucht – Ankyrin Repeat Proteine und Armadillo Repeat Proteine. Das Ankyrin Repeat Motiv ist ein weit verbreitetes Proteinsequenzmotiv und kommt in allen drei *Regna* vor. Ihre wichtigste Eigenschaft ist ihre Fähigkeit dreidimensionale Epitope anderer Proteine zu binden und dadurch mit ihnen zu interagieren. Armadillo Repeat Proteine sind üblicherweise an Protein-Protein oder Protein-Peptid Interaktionen beteiligt und können Peptide oder ungefaltete Teile von Proteinen binden. Ihre langgestreckte Bindefläche bindet Peptide in gestreckter Konformation.

Protein Engineering hat zum Ziel nützliche Proteine mit neuen oder verbesserten Eigenschaften zu entwickeln. Die Plückthun Gruppe hat in einem weitreichenden Design Projekt ein äusserst stabiles Consensus Ankyrin Repeat Proteingerüst (DARPins) entwickelt. Dieses optimierte Proteingerüst, besitzt spezifische Oberflächenreste deren Aminosäuren ausgetauscht werden können um Bindungsspezifität für ein beliebiges Protein herzustellen, ohne die Stabilität des Proteingerüsts negativ zu beeinflussen.
Repeat Proteine und insbesondere DARPins sind ein interessantes Modell um Proteinfaltung zu untersuchen, um die molekularen Hintergründe ihrer überraschend hohen Stabilität zu verstehen. An diesen designten Proteinen kann man in einer modularen und einheitlichen Umgebung Mechanismen der Proteinfaltung und –stabilität untersuchen.

In dieser Arbeit wurde die Stabilität und das Faltungsverhalten von DARPins, die aus Konsensusmodulen identischer Aminosäuresequenz aufgebaut sind, mittels NMR Spektroskopie, biophysikalischen und computerbasierten Methoden untersucht. Durch die identische Sequenz der Repeatmodule repräsentiert dieses Modell ein allgemeingültiges Beispiel für die Faltung von Ankyrin Repeat Proteinen und ermöglicht Untersuchungen ihres Faltungsverhaltens in Abhängigkeit von der Anzahl der vorhandenen Repeatmodule.
Mit Hilfe von Protonenaustauschmethoden wurde die Stabilität von Ankyrin Repeat Proteinen mit hoher Auflösung für die Amidprotonen des Proteinrückgrats unter nativen Bedingungen und in Gegenwart chemischer Denaturierungsmittel untersucht. Die Voraussetzung für diese Analyse war eine vollständige Zuordnung der Signale des

Proteinrückgrats der internen und der N- und C-terminalen Repeatmodule. Diese Aufgabe wird umso komplizierter je mehr identische interne Repeatmodule das jeweilige Protein aufweist. Paramagnetische Spinlabel welche an den N- und C-terminalen Repeatmodulen angebracht wurden, wurden eingesetzt um Signale identischer Positionen in verschiedenen Repeatmodulen zu unterscheiden und ermöglichten eine vollständige Zuordnung des Proteinrückgrats.

Unsere Ergebnisse zeigen, dass die Stabilität dieser DARPins sehr stark durch die Interaktionen zwischen den einzelnen Repeatmodulen beeinflusst wird. Einige Amidprotonen benötigten mehr als ein Jahr bei 37°C für einen vollständigen Austausch, was die immense Stabilität dieser Proteine hervorhebt. Faltungs- und Stabilitätsdaten wurde mit Vorhersagen eines Ising-basierten Modells verglichen. Das Entfaltungsverhalten unter Einfluss chemischer Denaturierungsmittel, welches mittels Protonenaustausch, Veränderungen der chemischen Verschiebung und heteronukleärem Overhausereffekt verfolgt wurde, stimmte mit dem Ising-Modell der Gleichgewichtsfaltung überein. In diesem Modell hängt die Stabilität jedes einzelnen Repeatmoduls von der Anzahl der Nachbarmodule ab und ist somit am höchsten für Repeatmodule die sich im Zentrum des Proteins befinden. Für den Austausch unter nativen Bedingungen postulieren wir, dass lokale Fluktuationen die Austauschraten dominieren, da Entfaltungsereignisse in diesen stabilen Proteinen zu langsam stattfinden. Die Lage der am langsamsten austauschenden Protonen deutet auf ein sehr stabiles hydrophobes Proteinzentrum in DARPins hin, in welchem hydrophobe Abschirmung und günstige elektrostatische Wechelwirkungen zusammenwirken. Diese Untersuchungen ermöglichen ein besseres Verständnis der Repeat Protein Architektur und erleichtern künftige Entwicklungen von DARPins für biomedizinische Anwendungen bei denen eine hohe Stabilität entscheidend ist.

Wie zuvor erwähnt sind Armadillo Repeat Proteine auf Grund ihrer Fähigkeit Peptide in gestreckter Konformation zu binden von besonderem Interesse. Dadurch können sie die Sequenz des Peptids anstatt, wie die meisten Proteinbinder, nur die dreidimensionale Struktur des Liganden erkennen. Der modulare Aufbau von Repeat Proteinen stellt eine gute Grundlage für die Entwicklung eines Gerüsts zur sequenzspezifischen Bindung von Peptiden dar. Ein solches Proteingerüst könnte durch beliebiges Aneinanderreihen und Vertauschen von Repeatmodulen an die Sequenz und Länge des zu bindenden Peptides angepasst werden. Die Grundlage dieser Designstrategie, auf der die in dieser Arbeit beschriebenen Projekte beruhen, bildet der konservierte Modus mit welchem das Peptidrückgrat von Peptiden durch natürliche Armadillo Repeat Proteine gebunden wird. Jedes Repeatmodul kann eine Dipeptideinheit über eine konservierte polare “Leiter”, bestehend meist aus Asparaginen, binden. Weitere Wechselwirkungen der Peptidseitenketten mit der Proteinoberfläche tragen zu den hohen Bindungsaffinitäten in natürlichen Peptid-Armadillo Repeat Proteinkomplexen bei.
Anfängliche Designs von auf Konsensussequenzen basierenden Armadillo Repeat Proteinen zeigten unzureichende Stabilität. Aufbauend auf Konsensus Armadillo Repeat Proteinen welche von Parmeggiani *et al.* Entwickelt wurden, wurden Untersuchungen durchgeführt um ein stabileres Proteingerüst zu entwickeln, dass sich für die Selektion spezifischer Peptidbinder eignet. Mit Hilfe von Molekulardynamiksimulationen, unterstützt durch NMR Experimente und verschiedene biophysikalische Methoden konnte sowohl die thermodynamische Stabilität der N- und C-terminalen Repeatmodule, als auch die der internen M-Typ Module verbessert werden. Heteronukleäre NMR Experimente zeigten ungünstige Interaktionen bei neutralem pH-Wert auf, welche bei

hohem pH keinen Effekt hatten. Desweiteren wurden Mutationen, die durch Simulationen vorgeschlagen worden waren, mit NMR Spektroskopie und temperaturabhängigen und chemischen Denaturierungsstudien untersucht. Die hierdurch entwickelten Proteine zeigten monomeres, kooperatives Faltungsverhalten und erhöhte Stabilität gegenüber Temperatur und chemischen Denaturierungsmitteln. Diese Ergebisse zeigen, dass Molekulardynamiksimulationen in Kombination mit NMR Spektroskopie und biophysikalischen Methoden eine gute Basis für die erfolgreiche Selektionierung von Mutationen darstellt, mit der biophysikalische Eigenschaften von Proteinen effizient verbessert werden können.

Die Zuordnung chemischer Verschiebungen von Repeat Proteinen durch NMR Methoden stellt wegen der sich wiederholenden Sequenzmotiven eine grosse Herausforderung dar. Um diesen Prozess zu erleichtern wurde versucht Proteine mit Hilfe einer Split-Intein Strategie partiell isotopenspezifisch zu markieren. Wir beobachteten, dass die einzeln exprimierten Proteinfragmente in Anwesenheit der Inteinmotive miteinander interagierten, jedoch ohne eine kovalente Bindung zu bilden. Dies konnte auch in Abwesenheit der Inteinmotive beobachtet werden, was darauf hindeutet, dass diese Wechselwirkung nicht auf Grund des Split-Inteins auftritt. Hierdurch entdeckten wir die bisher unbekannte Eigenschaft von Armadillo Repeat Proteinfragmenten spontan Komplexe zu bilden. Wir halten diese Erkenntnis für nützlich um die Zuordnung chemischer Verschiebungen in komplizierten Repeat Proteinen zu erleichtern. Um dies nachzuweisen wurden strukturelle, biophysikalische und thermodynamische Studien dieser Interaktion am Beispiel der Fragmente eines designten Konsensus Armadillo Repeat Proteins mit drei internen Repeatmodulen durchgeführt. Die Interaktion der beiden Fragmente wurde in Bezug auf die Gesamtstruktur des Komplexes und das Verhalten der einzelnen Fragmente untersucht. Wir konnten zeigen, dass der Fragmentkomplex dieselben Kontakte in der Interaktionsfläche zwischen Repeatmodulen ausweist wie das kovalent verbundene Ursprungsprotein. Diese Eigenschaft von Armadillo Repeat Proteinen könnte für fragmentbasiertes Design von Peptidbindern hilfreich sein, welches den effizienten Austausch von nicht kovalent gebundenen Fragmenten erlauben würde.

Das Hauptprojekt dieser Arbeit stellte die Charakterisierung der Interaktion eines von Varadamsetty *et al.* etablierten selektionierten Peptidbinders VG_328 mit dem Neurotensin Peptid dar. Wiederum stellte Signalüberlappung auf Grund von repetitiven Sequenzen in diesem Protein mit fünf internen Repeatmodulen eine grosse technische Herausforderung für die detaillierte Analyse mit NMR Spektroskopie dar. Nur ein geringer Anteil des Proteinrückgrates konnte mit klassischen dreidimensionalen NMR Experimenten zugeordnet werden. Zusätzliche Signalzuordnungen konnten durch die Analyse einer Reihe von verkürzten Proteinfragmenten erzielt werden. Obwohl diese Zuordnungen nicht vollständig waren, ermöglichten sie die Identifizierung von Repeatmodulen, welche nicht an der Bindung des Peptids beteiligt waren, basierend auf einer Analyse der Veränderungen chemischer Verschiebungen bei Peptidzugabe. In Folge dieser Analyse wurde Repeatmodule, die nicht für die Bindung von Neurotensin verantwortlich waren aus dem Protein entfernt, was zur Etablierung einer verkürzten Version von VG_328 führte, die vergleichbare Affinität für Neurotensin aufweist. Für dieses minimierte Binderprotein konnte das Proteinrückgrat vollständig zugeordnet werden. Daten aus Experimenten mit paramagnetisch markiertem Neurotensin, Mutationsstudien und Bindungsaffinitätanalysen wurden mit Vorhersagen des Bindungsmodus aus Molekulardynamiksimulationen verglichen und ein

übereinstimmendes Bild für die Konformation von Neurotensin identifiziert. Im diesem Bindungsmodell bindet der mittlere Teil von Neurotensin (Peptidreste 7-11) antiparallel zum Verlauf der Proteinsequenz entlang der Oberfläche welche von den dritten Helices der internen Repeatmodule gebildet wird. Diese Interaktion mit dem oberen Teil der designten Bindungsfläche beinhaltet Wechselwirkungen mit zahlreichen randomisierten Positionen auf der Proteinoberfläche. Der N-terminale Teil des Peptids (Peptidreste 1-7) scheint sehr flexibel zu sein und kann eine Vielzahl verschiedener Konformationen einnehmen. Der C-Terminus des Peptids hingegen kann eine stabile Salzbrücke mit einer Argininseitenkette der zweiten Helix des N-terminalen Repeatmoduls des Proteins bilden und dadurch den Protein-Peptidkomplex stabilisieren. Diese Erkenntnisse wurden durch Mutationsstudien des N-terminalen Repeatmoduls weiter untermauert, welche Einblicke in die Bedeutung der Mutationen auf die Proteinstabilität und die Peptidbindung gewährten. Diese Studien zeigten den detaillierten Einfluss einer kleinen Gruppe von Mutationen auf die Faltung des N-terminalen Repeatmoduls und dessen Kontakt mit dem ersten internen Repeatmodul und wiesen auf ein empfindliches Gleichgewicht zwischen Proteinstabilität und Peptidbindung in diesem Peptid-Proteinkomplex hin.

Unsere Strategie zeigte, dass NMR Daten niedriger Auflösung ausreichend waren, um weitere Designschritte einzuleiten, und dass eine Kombination aus NMR und Molekulardynamikdaten den Bindungsmodus des Peptids ohne kristallografische Daten für diesen schwachen Liganden erfolgreich bestimmen konnten. Dies ist von besonderer Bedeutung in der Anfangsphase von Protein Designprojekten, wenn Bindungsaffinitäten noch sehr schwach sind und nicht optimal stabilisierte Proteine für kristallografische Studien ungeeignet sind. Wir haben in diesem Projekt eine stark interdisziplinäre Strategie in der NMR Spektroskopie, Molekulardynamiksimulationen und andere biophysikalische Methoden kombiniert wurden entwickelt. Wir halten diese Strategie für vielversprechend um komplizierte Systeme zu untersuchen, wie zum Beispiel Bindungskomplexe mit niedriger Affinität die mehrere Ligandenkonformationen und suboptimale Stabilität aufweisen. Desweiteren konnten wir zeigen, dass begrenzte Signalzuordnungen durch NMR Daten ausreichen um Proteinsequenzen so anzupassen, dass neue Proteine mit verbesserten Eigenschaften entwickelt werden können, die für hochauflösende strukturelle Studien geeignet sind. Diese begrenzten Daten waren ausreichend um das Projekt voranzutreiben und um ursprüngliche Hypothesen des Bindungsmodus des Liganden an das designte Bindeprotein zu überprüfen. Besonders zu Beginn von Designprojekten ist es wichtig zu zeigen, dass das Projekt auf dem korrekten Weg ist, um gegebenenfalls Anpassungen vornehmen zu können.

Table of Contents

Acknowledgements iv
Summary v
Zusammenfassung ix
List of Figures Chapter 1 xviii
List of Figures Chapter 2 xviii
List of Supplementary Figures Chapter 2 xix
List of Figures Chapter 3 xix
List of Supplementary Figures Chapter 3 xx
List of Figures Chapter 4 xx
List of Supplementary Figures Chapter 4 xx
List of Figues Chapter 5 xxi
List of Tables xxii
1. Introduction 1
1.1 Repeat proteins 1
1.1.1 Overview 1
1.1.2 Ankyrin Repeat Proteins 4
1.1.3 Armadillo Repeat Proteins 5
1.2 Protein Engineering and Consensus Design of Repeat Proteins 7
1.2.1 Overview of Protein Engineering 7
1.2.2 Rational Protein Design 7
1.2.3 Directed Evolution and Ribosome Display 8
1.2.4 The Design and Applications of the DARPin Scaffold 11
1.2.5 Scaffold Choice for a Modular Peptide Binding Platform 11
1.2.6 Ankyrin Repeat Protein Folding 21
1.2.7 The Ising Model in Repeat Protein Folding 23
1.3 NMR Spectroscopy of Proteins 24
1.3.1 Introduction 24
1.3.2 Heteronuclear Single Quantum Coherence - The HSQC 24
1.3.3 Resonance Assignment of Proteins 25
1.3.4 NMR of Repeat Proteins 27
1.3.5 Protein Stability Determination by NMR Spectroscopy 28

1.3.6 Probing Weak Protein Ligand Interactions 29
1.4 Project Goals 31
1.5 References 32

2. Residue-resolved stability of full-consensus ankyrin repeat proteins probed by NMR 40
2.1 Abstract 41
2.2 Introduction 42
2.3 Results 44
2.3.1 Backbone assignment of NI2C, NI3C and NI3C_Mut5 44
2.3.2 Residue-resolved stability mapping using amide proton exchange 47
2.3.3 Equilibrium denaturant unfolding of NI_3C and NI_3C_Mut5 analyzed by NMR 52
2.3.4 Comparison to calculations based on the Ising model 55
2.4 Discussion 58
2.4.1 Stability-determining role of the C-cap 58
2.4.2 Coupling of adjacent repeats largely influences the folding energy landscape 59
2.4.3 Stabilizing features of the DARPins 61
2.4.4 Comparison with other repeat proteins 62
2.5 Conclusions 63
2.6 Materials and Methods 64
2.6.1 Protein biochemistry and production of spin-labeled proteins 64
2.6.2 Production of spin-labeled proteins: 65
2.6.3 Spin-label experiments: 65
2.6.4 NMR Spectroscopy and Data Evaluation 66
2.6.5 Measurement of amide proton exchange 67
2.6.6 Data analysis of amide proton exchange 67
2.6.7 Measurement of GdmCl or pH-induced equilibrium unfolding by NMR . 68
2.6.8 Fit to Ising model 68
2.7 Acknowledgements 69
2.8 References: 70
2.9 Supplementary Material for Wetzel *et al.* 73
2.9.1 Some remarks considering the assignment: 73

2.9.2 Supplementary Figures 74

3. Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy 90

3.1 Abstract 91

3.2 List of Abbreviations: 92

3.3 Introduction 93

3.4 Results 95

3.4.1 Optimization of the internal repeats using heteronuclear NMR spectroscopy 95

3.4.2 MD simulations suggest mutations at the N-cap and C-cap that result in improved protein stability 97

3.4.3 Biophysical characterization of the M- and M̄-type proteins 104

3.4.4 Biophysical characterization of various cap mutants allows identifying mutants with much improved stability 107

3.4.5 Biophysical characterization of cap combinations 109

3.5 Discussion 112

3.6 Materials and Methods 113

3.6.1 Nomenclature 113

3.6.2 MD simulations 114

3.6.3 Clustering of trajectories 114

3.6.4 Trajectory analysis 115

3.6.5 Model generation 115

3.6.6 Design and synthesis of DNA encoding designed ArmRPs, protein expression and purification 115

3.6.7 Protein purification 116

3.6.8 Circular dichroism spectroscopy 116

3.6.9 ANS fluorescence spectroscopy 116

3.6.10 Size exclusion chromatography and multi-angle light scattering 117

3.6.11 NMR Spectroscopy 117

3.7 References 118

3.8 Supplementary Material to 121

3.8.1 Supplementary Materials and Methods 121

4. Spontaneous Self-Assembly of Fragments of Engineered Armadillo Repeat Proteins into Folded Proteins 129
4.1 Introduction 130
4.2 Results and Discussion 131
4.2.1 Properties of the fragments 131
4.2.2 Properties of the complex 133
4.3 Conclusions 137
4.4 Acknowledgements 138
4.5 References: 138
4.6 Supplementary Materials 140
4.6.1 Methods and Materials 140
4.6.2 Supplementary References 158

5. An Interdisciplinary Approach to Investigate Peptide Binding to a Designed Armadillo Repeat Protein Improving Protein Design using NMR, MD and other Biophysical Techniques 159
5.1. Abstract 160
5.2. Introduction 161
5.3. Materials & Methods 166
5.3.1. Nomenclature 166
5.3.2. Molecular Biology 166
5.3.3. Cloning of $Y_IMR^1R^2R^3MA_{II}$ Fragments 167
5.3.4. Cloning of $Y_IMR^1R^2R^3A_{II,}$ $Y_IMR^1R^2A_{II}$ and $Y_IMR^1A_{II}$ 168
5.3.5. Cloning of $Y_IMR^1R^2A_{II}$-Mutants 168
5.3.6. Expression of Unlabeled and Isotopically Labeled Proteins 168
5.3.7. Protein Purification and Characterization 169
5.3.8. NMR Spectroscopy and Data Evaluation 170
5.3.9. Backbone and Side Chain Assignment 170
5.3.10. Chemical Shift Mapping (CSM) Experiments 171
5.3.11. Determination of Dissociation Constants (K_d) by [^{15}N,^{1}H]-HSQC based CSM Titrations 171
5.3.12. Determination of Dissociation Constants (K_d) by Surface Plasmon Resonance (SPR) 172
5.3.13. Paramagnetic Relaxation Enhancement (PRE) Experiments 173

5.3.14. ELISA Assays 173
5.3.15. Molecular Dynamics Simulations 174
5.4. Results 175
5.4.1. Overview 175
5.4.2. Stabilization of VG_328 for NMR Studies 176
5.4.3. Truncation of $Y_IMR^1R^2R^3MA_{II}$ Aids in Backbone Assignment and Reveals Contributions of Individual Repeats to Protein Stability 177
5.4.4. Design of the Minimal-Size Binder $Y_IMR^1R^2A_{II}$ Reduces Target Complexity 181
5.4.5. Interaction Studies of $Y_IMR^1R^2A_{II}$ and NT Using CSP and PRE Data, and MD Simulations 183
5.4.6. Binding Strengths of NT Towards Various N-cap Mutants of $Y_IMR^1R^2A_{II}$ as Probed by NMR 189
5.4.7. Binding Strengths of NT Towards Various N-cap Mutants of $Y_IMR^1R^2A_{II}$ Confirmed by ELISA and SPR Studies 192
5.4.8. Probing Protein Stability with MD Simulation 194
5.4.9. Identifying Potential Binding Conformations of NT Using MD Simulations of $Y_IMR^1R^2A_{II}$ and its Variants in Complex with NT 197
5.5. Discussion and Conclusions 201
5.6. Supplementary Materials 206

List of Figures Chapter 1

Figure 1.1 Representative structures of six main repeat protein families..... 3
Figure 1.2 Details of ankyrin repeat proteins..... 5
Figure 1.3 Structural details of armadillo repeat proteins for yeast karyopherin-α 6
Figure 1.4 The principals of ribosome display. 10
Figure 1.5 Peptide binding by an antibody. 12
Figure 1.6 The peptide binding modes of four major repeat protein families.. 14
Figure 1.7 Details of the binding mode of importin-α. 16
Figure 1.8 Comparison of peptide binding modes between antibodies and the armadillo repeat proteins importin-α and β-catenin.. 17
Figure 1.9 Schematic illustrating the strategic combination of library selection and repeat shuffling. 18
Figure 1.10 Sequence alignment of designed armadillo repeat proteins. 19
Figure 1.11 Schematic contributions to the free energy of unfolding according to the Ising model of protein folding. 23
Figure 1.12 Illustrative summary of the main 3D NMR data sets utilised for the resonance assignment of backbone nuclei.. 26
Figure 1.13 Correlations from 3D NMR experiments used to determine side chain resonance assignments.. 27
Figure 1.14 Chemical structures of the spin labels MTSL and NHS..... 30

List of Figures Chapter 2

Figure 2.1 700 MHz [^{15}N, ^{1}H]- HSQC spectrum of 1.5 mM ^{15}N,^{13}C,^{2}H-labeled NI_3C_Mut5 in 50 mM phosphate, 150 mM NaCl, pH 7.4 at 310 K. 44
Figure 2.2 Paramagnetic relaxation enhancement (PRE) data of D28C and D155C mutants of NI_3C_Mut5. 46
Figure 2.3 ^{15}N{^{1}H}-NOE data for NI_2C recorded at 600 MHz. 47
Figure 2.5 Sample signal decay curves of NI3C in hydrogen exchange experiments. 48
Figure 2.6 Hydrogen exchange data for NI_2C, NI_3C and NI_3C_Mut5..... 49
Figure 2.7 Mapping of the protection factors by colour 50
Figure 2.8 Expansion from 700 MHz [^{15}N,^{1}H]-HSQC spectra of NI_3C at GdmCl concentrations of 0, 0.6, 1.2, 1.8 and 2.4 M..... 53
Figure 2.9 600 MHz [^{15}N,^{1}H]-HSQC spectra of NI_3C_Mut5 in presence of 4 M GdmCl.. 55
Figure 2.10 Protection factors for selected, slowly exchanging residues of the internal repeats I-1, I-2 and I-3 of NI_3C_Mut5 in native buffer or 3.5 M GdmCl... 56
Figure 2.11 Expectations of the protection factors at 293 K derived from an Ising-type folding model for NI_2C, NI_3C and NI_3C_Mut5 57

List of Supplementary Figures Chapter 2

Figure S2.1 Sequence alignment of the full-consensus ankyrin repeat proteins NI_2C, NI_3C and the C-cap mutant NI_3C_Mut 5. 74
Figure S2.2 600 MHz [^{15}N,^{1}H]-HSQC spectrum of NI_2C 75
Figure S2.3 600 MHz [^{15}N,^{1}H]-HSQC spectrum of NI_3C 76
Figure S2.4 600 MHz [^{15}N,^{1}H]-HSQC spectrum of NI_3C_Mut5 77
Figure S2.5 Representative strips from the 3D HNCACB and HN(COCA)NH spectra of NI_3C_Mut5 78
Figure S2.6 *Intramolecular* attenuations of signal intensities of cross peaks in the 600 MHz [^{15}N,^{1}H]-HSQC spectra of MTSL-derivatized D28C-NI_2C and D28C-NI_3C 79
Figure S2.7 *Intermolecular* attenuations of signal intensities of cross peaks in the 600 MHz [^{15}N,^{1}H]-HSQC spectra of a mixture of MTSL-derivatized unlabelled D28C-NI_3C, D28C-NI_3C_Mut5 or D155C- NI_3C_Mut5 with the non-spin labeled ^{15}N uniformly labeled corresponding proteins protected wit NEM. 80
Figure S2.8 700 MHz [^{15}N,^{1}H]-HSQC spectra of NI_3C at 0 M, 1.2 M, 2.4 M and 3.6 M GdmCl 81
Figure S2.9 ^{1}H and ^{15}N chemical shift changes mapped onto the structure of NI_3C and NI_3C_Mut5. 82
Figure S2.10 ^{15}N-{^{1}H}-NOE data of NI_3C_Mut5 in presence of 2 M, 4 M or 6 M GdmCl recorded at 600 MHz. 83
Figure S2.11 700 MHz [^{15}N,^{1}H]-HSQC spectra of NI_3C_Mut5 at 0 M, 2.0 M, 4.0 and 6.0 M GdmCl 84
Figure S2.12 ^{15}N-{^{1}H}-NOE data of NI_2C, NI_3C and NI_3C_Mut5 85
Figure S2.13 NI_3C in presence of 1 M and 2 M GdmCl 86
Figure S2.14 CD-monitored denaturation curves of NI_1C, NI_2C, NI_3C, NI_1C_Mut5 and NI_3C_Mut5 and the Ising model fit 87
Figure S2.15 600 MHz [^{15}N,^{1}H]-HSQC spectrum of NI_3, 310 K 88
Figure S2.16 Predictions of α-helix content for NI_3C_Mut5 based on the primary sequence using the program AGADIR 89

List of Figures Chapter 3

Figure 3.1 An armadillo repeat protein bound to a peptide. Importin-α in complex with a nucleoplasmin NLS peptide. 94
Figure 3.2 Representative [^{15}N,^{1}H]-HSQC spectra of YM_4A recorded at various values of pH 96
Figure 3.3 $Y_{II}\bar{M}_4A_{II}$ (QQ-type) model displaying the location of stabilizing mutations as sticks 97
Figure 3.4 Water molecules permeate into the R_4/C interface of YM_4A 98
Figure 3.5 Analysis of implicit solvent MD simulations of YM_4A variants. 99
Figure 3.6 Analysis of explicit water MD simulations of YM_4A variants 102
Figure 3.7 Distance distribution of the salt bridge between the N-cap and the first repeat 103
Figure 3.8 Biophysical characterization of designed ArmRP 105
Figure 3.9 Biophysical characterization of designed ArmRPs $Y\bar{M}_4A$ and its cap variants 106
Figure 3.10 [^{15}N,^{1}H]-HSQC spectra of designed ArmRP $Y\bar{M}_3A$ and its cap variants. 111

List of Supplementary Figures Chapter 3

Figure S3.1 Sequence of the armadillo repeats and mutants studied. 124
Figure S3.2 RMSF plot of Cα atoms of YM_4A in the implicit solvent simulations. 124
Figure S3.3 RMSF plot superposition of all implicit solvent simulations 125
Figure S3.4 RMSF plot superposition of all explicit solvent simulations 125
Figure S3.5 Mutations introduced into the N_{II}-cap compared to the wild-type. 126
Figure S3.6 Single-step IMAC purification of YM/$\bar{M}_4$A N-and C-cap variants. 126
Figure S3.7 Biophysical characterization of designed ArmRPs $Y\bar{M}_3A$ and its cap variants $Y_{II}\bar{M}_3A$, $Y\bar{M}_3A_{II}$ and $Y_{II}\bar{M}_3A_{II}$. 127
Figure S3.8 [^{15}N,^{1}H]-HSQC spectra of designed ArmRP $Y\bar{M}_4A$ and its cap variants 128

List of Figures Chapter 4

Figure 4.1 Amino acid sequences of the two fragments investigated in this study 131
Figure 4.2 [^{15}N, ^{1}H]-HSQC spectra of: ^{15}N MA and ^{15}N MA complexed with unlabeled YM_2, ^{15}N labeled YM_2 alone; and ^{15}N labeled YM_2 complexed with unlabeled MA 132
Figure 4.3 Isothermal titration calorimetry isotherm and curve fitting for the YM_2/MA interaction 134
Figure 4.4 A, Solution structure of uncomplexed MA displaying the ensemble of the 20 lowest-energy conformers. B, Solution structure of uncomplexed MA superimposed with the corresponding region from the crystal structure of the entire protein, YM_3A. C, Solution structure of MA complexed with YM_2. D, Ensemble of the 20 lowest-energy structures superimposed over the crystal structure of YM_3A. 135
Figure 4.5 Image indicating MA and YM_2 and a selection of observed inter-molecular NOEs, establishing the relative orientations of the two fragments. 136

List of Supplementary Figures Chapter 4

Figure S4.1 Schematic overview of expression products YM_2 and MA and 15% SDS-PAGE analysis 141
Figure S4.2 Preparative and analytical size exclusion analysis of YM_2, MA and their complex. 145
Figure S4.3 CD spectra of YM_2 and MA. 146
Figure S4.4 Thermal denaturation observed at 220 nm for YM_2 and MA 147
Figure S4.5 600 MHz [^{15}N,^{1}H]-HSQC spectrum of uncomplexed MA 148
Figure S4.6 600 MHz [^{15}N,^{1}H]-HSQC spectrum of MA complexed with 1.2 equiv. of YM_2. 149
Figure S4.7 700 MHz [^{15}N,^{1}H]-HSQC spectrum of YM_2 complexed with 1.5 equiv. of MA. 150
Figure S4.8 NOE restraints per residue as used in the final cycle for the calculation of the uncomplexed MA structure. 153
Figure S4.9 NOE restraints per residue as used in the final cycle for the calculation of the complexed MA structure. 154
Figure S 4.10 Chemical shift deviations (δ_{av}) for MA upon complex formation with NMR-invisible YM_2. 155
Figure S4.11 Probability (P) for formation of helical or coiled dihedrals using the program TALOS+ 156

List of Figues Chapter 5

Figure 5.1 General strategy for NMR assignment and characterisation of NT peptide binding to an engineered armadillo repeat protein based on VG_328. 163
Figure 5.2 The sequence of proteins based on $Y_IMR^1R^2R^3MA_{II}$ and associated modifications and mutations.. 164
Figure 5.3 [^{15}N,^{1}H]-HSQC spectra of 300 µM N-terminally truncated $Y_IMR^1R^2R^3MA_{II}$ fragments. 178
Figure 5.4 The assignment strategy using successive addition of repeat modules utilising the spectra of the N-terminally truncated fragments. Glycine residues in an overlay of [^{15}N,^{1}H]-HSQC spectra of 300 µM N-terminally truncated $Y_IMR^1R^2R^3MA_{II}$ fragments. 179
Figure 5.5 Chemical shift perturbations in [^{15}N,^{1}H]-HSQC spectra of 400 µM ^{15}N-labeled $Y_IMR^1R^2R^3MA_{II}$ without and with two equivalents NT. 180
Figure 5.6 Chemical shift perturbations of $Y_IMR_1R_2R_3MA_{II}$ upon addition of two equivalents of NT mapped onto a structural model. 181
Figure 5.7 Expansion of the Gly region in [^{15}N,^{1}H]-HSQC spectra of 400 µM ^{15}N-labeled $Y_IMR^1R^2R^3MA_{II}$, $Y_IMR^1R^2R^3A_{II}$, $Y_IMR^1R^2A_{II}$, $Y_IMR^1MA_{II}$ without and in complex with two equivalents NT. 182
Figure 5.8 CSPs for $Y_IMR^1R^2A_{II}$ upon NT binding mapped onto a structural model based on the unrandomized designed armadillo repeat protein $Y_{III}M_3A_{II}$ 184
Figure 5.9 Top: Signal attenuation observed in PRE experiments of $Y_IMR^1R^2A_{II}$ with spin labelled NT peptides mapped onto a structural model and list of peptides used for CSP and PRE studies. 186
Figure 5.10 Comparison of $Y_IMR^1R^2A_{II}$ and $Y_{III}M_3A_{III}$ analyzing the rotational movement of the N-cap of $Y_IMR^1R^2A_{II}$. 188
Figure 5.11 Dissociation constants (K_d) as determined by NMR using CSPs. 190
Figure 5.12 Fitted CSP raw data of exemplary $Y_IMR^1R^2A_{II}$ residues 191
Figure 5.13 Affinity of $Y_IMR^1R^2A_{II}$ mutants and VG_328 based reference proteins for NT1-13 detected by ELISA.. 193
Figure 5.14 SPR response curves and fitted data for $Y_IMR^1R^2A_{II}$ and $Y_IMR^1R^2R^3MA_{II}$ 194
Figure 5.15 Comparison of RMSF values for the N-cap and the first internal repeat of selected $Y_IMR^1R^2A_{II}$ variants in presence and absence of NT. 195
Figure 5.16 Evolution of atomic distances in $Y_IMR^1R^2A_{II}$_V34R over the course of the trajectory. 196
Figure 5.17 Evolution of atomic distances in $Y_IMR^1R^2A_{II}$_V34R_R37S during the MD trajectory. 197
Figure 5.18 Peptide conformation picked at 19.465 µs. $Y_IMR^1R^2A_{II}$ is displayed in blue, NT in yellow. 198
Figure 5.19 RMSD fluctuations of the peptide Cα-atoms over the trajectory in comparison to the reference peptide conformation picked at 19.465 µs 199
Figure 5.20 [^{15}N,^{1}H]-HSQC spectra showing chemical shift perturbations of 400 µM ^{15}N-labelled $Y_IMR^1R^2R^3MA_{II}$ without and with two equivalents of NT peptide. 208
Figure 5.21 [^{15}N,^{1}H]-HSQC spectra showing chemical shift perturbations of 400 µM ^{15}N labelled $Y_IMR^1R^2A_{II}$ without and with two equivalents of NT peptide. 209

List of Tables

Chapter 3:

Table 3.1 Mutants investigated by MD simulations 101
Table 3.2 Biophysical properties of designed ArmRPs with different capping repeats. 108

Table S3.1 List of oligonucleotides used for generating point mutants 123

Chapter 4:

Table S4.1 Sequences of Oligonucleotide Primers.....140
Table S4.2 Molecular weights of unlabeled fragments 141
Table S4.3 Composition of minimal medium.....142
Table S4.4 Composition of trace metal solution.....142
Table S4.5 Comparison of sequence based molecular weight and SEC-observed size.....145
Table S4.6 Structure statistics for uncomplexed MA 151
Table S4.7 Structure statistics for complexed MA 152
Table S4.8 Table of observed interface NOEs originating from MA, observed on YM_2.....157

Chapter 5:

Table 5.1 List of dissociation constants of $Y_IMR^1R^2A_{II}$ mutants and VG_328 based reference proteins.....190
Table 5.2 Oligonucleotides 206
Table 5.3 Minimal medium for isotopic labeling. 207
Table 5.4 Trace metal solution for minimal media supplementation. 207
Table 5.5 Peptides, free termini unless indicated otherwise. 207
Table 5.6 List of chemical shifts for backbone and side chain assignments of ^{13}C, ^{15}N-labeled $Y_IMR^1R^2A_{II}$ in the presence of 2 equivalents NT. 210
Table 5.7 List of chemical shifts for backbone assignments of ^{2}H, ^{13}C, ^{15}N-labeled $Y_IMR^1R^2A_{II}$ in the presence of 2 equivalents NT. 214
Table 5.8 List of chemical shifts for backbone assignments of ^{2}H, ^{13}C, ^{15}N-labeled $Y_IMR^1R^2A_{II}$. 216
Table 5.9 List of chemical shifts for backbone assignments of ^{2}H, ^{13}C, ^{15}N-labeled $Y_IMR^1R^2\ R^3MA_{II}$ in the presence of 2 equivalents of NT. 218
Table 5.10 List of chemical shifts for backbone assignments of ^{2}H, ^{13}C, ^{15}N-labeled $Y_IMR^1R^2\ R^3MA_{II}$ 220

1. Introduction

1.1 Repeat proteins

1.1.1 Overview

Repeat proteins (RPs), also called tandem repeat proteins, are a class of proteins constituting almost 20 % of proteins encoded in the human genome. In this class, repetitive amino acid sequence motifs emerged throughout evolution as a result of intragenic duplication and recombination events [1]. As a result these proteins contain tandem arrays of highly similar structural units of 20-60 amino acids in length [2]. These units or repeats generally contain simple secondary structure elements, e.g. a helix-turn-helix motif. Several repeats stack against each other forming non-globular, elongated structures. Their hereby created long hydrophobic cores and their extensive solvent exposed surfaces determine topology and function of this special family of proteins [3,4].
Depending on the RP family, 3-12 repeating units per protein are most commonly observed, however, up to 36 repeated units have been reported for HEAT RPs [2]. Specialized repeats, so-called caps, flank the array of internal repeats N- and C-terminally improving solubility and stability of RPs.

Repeat proteins differ from globular proteins in several important characteristics [5]. Firstly, they commonly display an extended solenoid fold [6,7], even though there are examples of circular repeat proteins where the first and last repeat interact forming a closed fold [8]. Secondly RP folds are mainly stabilized by short-range interactions between residues close in sequence forming an extended hydrophobic core. The importance of long-range interactions for protein stability is greatly diminished compared to globular proteins [9,10].

Table 1.1 Repeat protein families and basic characteristics. Adapted from Grove *et al.* 2008.

Repeat protein family	Number of amino acid residues	Repeating structural motif	Number of repeats occurring in natural setting (*Most common)
Ankyrin	30	Helix-helix-loop (or β-hairpin) (H1, H2)	4-24, 6*
Armadillo	42	Three α-helices (H1, H2, H3)	6-15, 12*
HEAT	37-47	Two α-helices (A, B)	3-36
Leucine Rich Repeats (LRR)	20-29	β-strand-loop-helix	Up to 28
Tetratricopeptide repeats (TPR)	34	Helix-turn-helix (A, B)	3-16, 3*
WD40	40-60	Four-stranded (a-d) antiparallel β-sheet	3-16, 7-8*

On the sequence level, natural RPs are not always characterized by identical or near identical repeating amino acid sequences. In some cases sequence identity can be very low [1,11], but for many repeat proteins a consensus sequence can be established.
Designed RPs with repeats of identical sequence have been shown to exhibit more regular structures and higher stability than in nature [12].

Natural repeat proteins fulfil a plethora of important functions in cell biology like molecular recognition, cell adhesion and transport. Well known representatives of this protein class include armadillo RPs, ankyrin RPs, HEAT RPs, tetratricopeptide RPs, WD40 RPs and many more. A selection of repeat protein families and their properties is listed in Table 1.1 and Figure 1.1

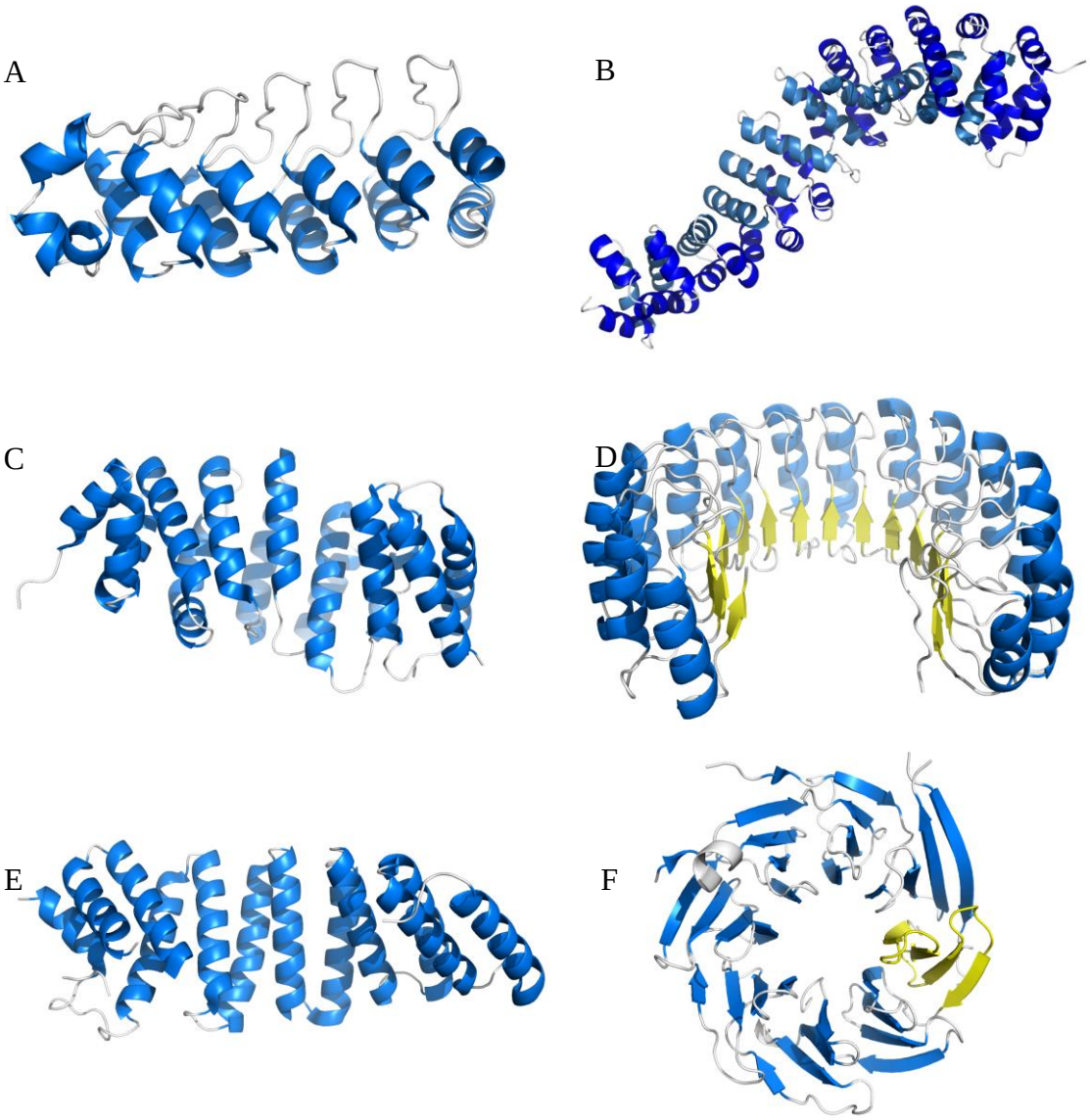


Figure 1.1 Representative structures of 6 main repeat protein families. A, Ankyrin repeat protein TV1425 from *Thermoplasma volcanicum* (PDB ID 2RFM); B, Armadillo repeat domain of murine β-catenin (PDB ID 2BCT); C, Archeal HEAT-like repeat protein TON_1937 from *Thermococcus omniurineus* NA1 (PDB ID 3B2A); D, Porcine ribonuclease inhibitor comprising leucine rich repeats (LLR) (PDB ID 2BNH); E, The tetratricopeptide repeat (TPR) domain of human kinesin light chain 2 (PDB ID 3CEQ); F, The C-terminal WD40 domain of Sif2 subunit of the Set3C histone deacetylase complex from *Saccharomyces cerevisiae* with repeating unit shown in yellow, (PDB ID 1R5M).

1.1.2 **Ankyrin Repeat Proteins**

Sequence analyses show that the natural ankyrin repeat (AR) is a very common type of motif and can be found in all three kingdoms [4,13]. AR proteins have been found spanning a wide range of functions, with the underlying theme being their ability to mediate protein-protein interactions.

Ankyrin repeats are present in about 6% of eurkaryotic protein sequences [14,15] and are involved in diverse functions including cell cycle regulation, signal transduction, transcription initiation and in cytoskeletal proteins. Destabilizing mutations in AR proteins contribute to several human diseases such as the case of the human erythrocyte protein AnkyrinR (originally called Band 2.1) which gave this repeat protein family its name [16]. In erythrocytes, AnkyrinR connects the spectrin of the membrane skeleton with an anion exchanger (Band 3) [17]. Reduced levels or defective AnkyrinR lead to mechanically weakened erythrocytes and are the basis of hereditary spherocytosis, a form of haemolytic anaemia [18].

The AR motif consists of 33 residues, which form a short β-turn followed by two α-helices before connecting to the following repeat with a longer loop [19] (see Figure 1.2 A) and was first identified in Cdc10, a cell cycle regulator in yeast, and the *Drosophila* signalling protein Notch[20].

Most AR proteins contain 4-7 repeats and are on average shorter than other repeat proteins such as armadillo repeat proteins, even though exceptions like AnkyrinR with 24 repeats are known.

In an array of ankyrin repeats flanked by protective caps, the loops and first helix of each repeat form an extended binding interface [21] (see Figure 1.2 B). The specificity of the elongated binding domain is determined by the varying surface residues [22] and can be altered by mutating these positions. The binding surface can easily be extended by adding more repeats to the array to provide space for larger interaction partners or create a multivalent binding site. Unlike other binding domains AR proteins are not limited to one specific target or binding pattern and are determined by their three-dimensional structure rather than function.

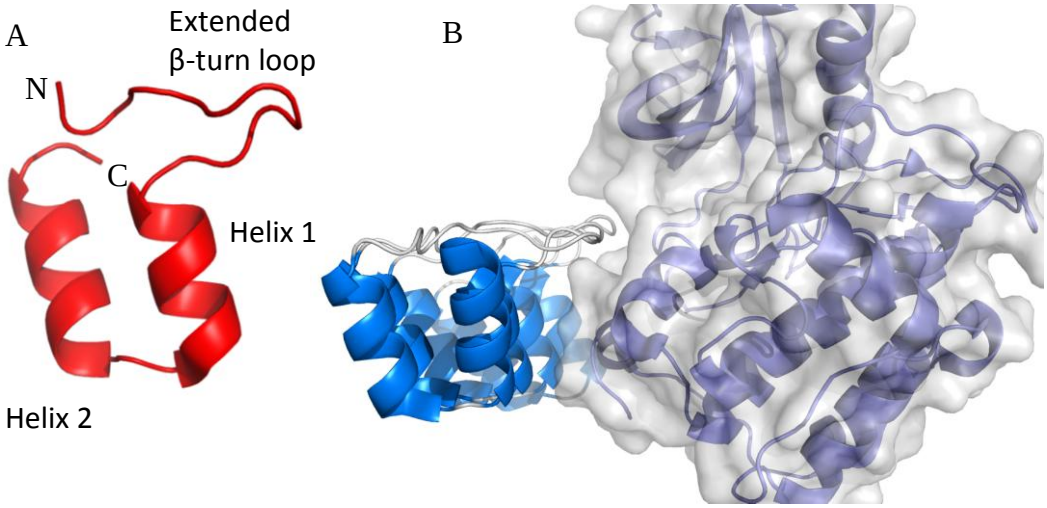


Figure 1.2 Details of ankyrin repeat proteins. A, The ankyrin repeat motif in red showing the two helices and the loop that forms an integral part of the binding interface; B, an designed ankyrin repeat protein (DARPin, light blue) selected to bind the wildtype PLK-1 Kinase domain (grey surface and dark blue cartoon), illustrating the binding of ankyrin repeat proteins to folded tertiary protein structure. (PDB ID 2V5Q).

1.1.3 **Armadillo Repeat Proteins**

Armadillo repeat proteins are commonly involved in protein-protein or protein-peptide interactions, binding to peptides or unfolded parts of proteins and are involved in regulating cell-cell adhesion, gene transcription, or nucleo-cytoplasmic transport.
The armadillo repeat motif typically consists of 42 amino acids folding into 3 α-helices termed helices 1-3. Helix 1 lies perpendicular to the hairpin between helices 2 and 3, which stack in an antiparallel manner. This results in a triangular structural motif [7] (see Figure 1.3 A). ArmRPs are typically found in longer arrays in nature (10-15 repeats) (see Table 1.1). Typically, 12 of these repeating triangles stack against each other to form a right-handed superhelix. The continuous surface formed by the array of all the third helices creates an extended binding surface (see Figure 1.3 B), which can bind peptides or unfolded parts of proteins in an extended conformation (see Figure 1.3C).

The name '*armadillo*' was first given to a segment polarity gene discovered by Christiane Nüsslein-Volhard and Eric Wieschaus in a *Drosophila melanogaster* mutant [23,24] in the early 1980s. The larvae of this mutant type displayed a distinct segmentation pattern caused by the duplication of the anterior part of each segment replacing the posterior part giving them a striped look akin to that of the armadillo animal. The human homolog of the *armadillo* gene is *β-catenin* involved in Wnt-signaling [25]. The term "armadillo repeat" was adopted by Riggleman *et al.* after sequence analysis showed the repetitive sequences in the *armadillo* gene and its expressed protein [26].

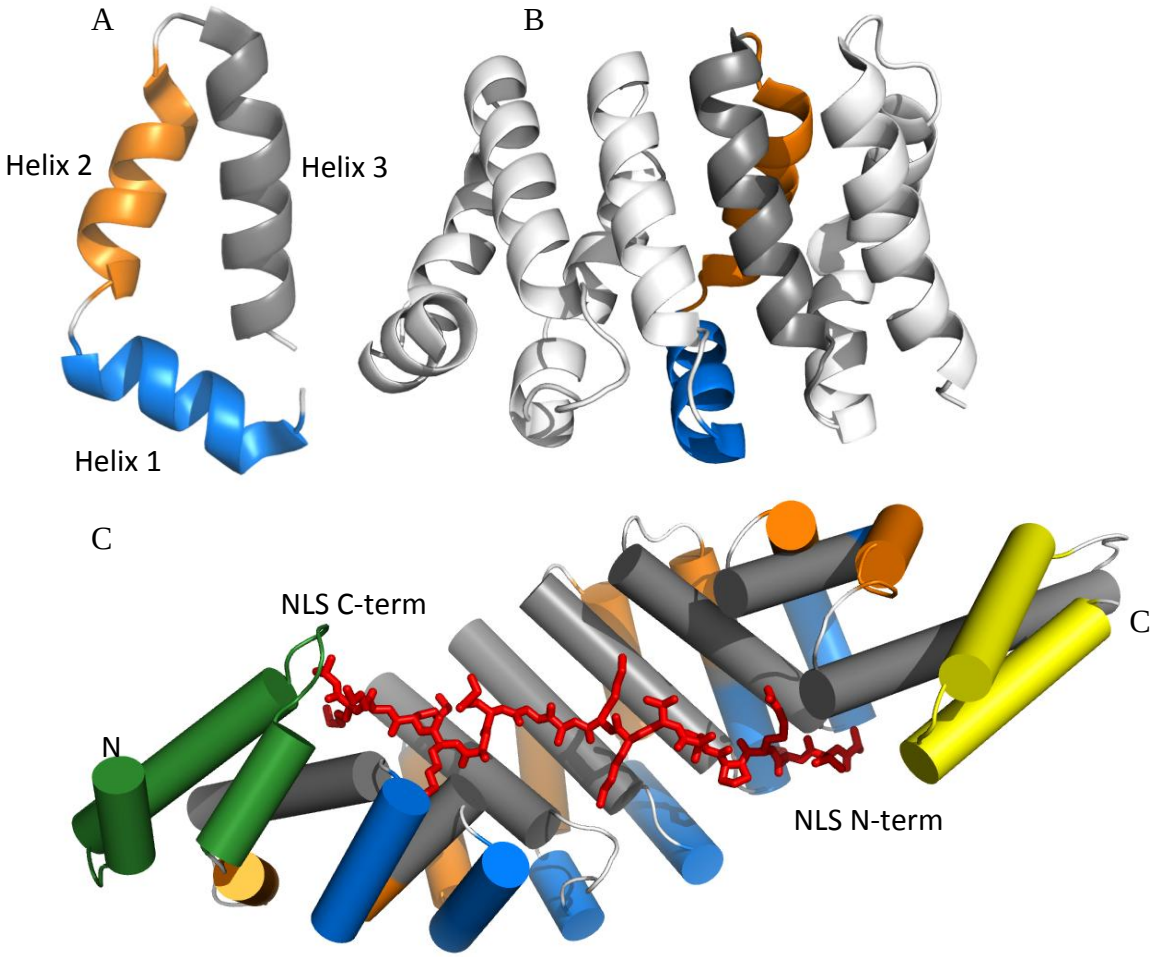


Figure 1.3 Structural details of armadillo repeat proteins depicted for yeast karyopherin- α (importin-α, PDB iD 1EE5). A, the armadillo repeat motif showing helix 1 (blue), 2 (orange), and 3 (grey); B, the packing of armadillo repeats in the context of multiple repeats, showing repeats 3-6; C, the binding of the nuclear localisation signal peptide (NLS, red sticks) in an anti-parallel manner and extended conformation across the interface formed by the adjacent 3rd helices of each repeat (grey), the N-termimus of the protein is depeicted in green, the C-terminus in yellow.

Peifer *et al*. established the first consensus sequence and proteins containing similar repeating sequences where thereafter categorized as armadillo repeat proteins [27].
The first structures of armadillo repeat proteins, mouse β-catenin[28] and yeast karyopherin-α [29] (human analog: importin-α) revealed the solenoid topology of the armadillo domain and were solved in the late 1990s by Huber *et al*. and Conti *et al*., respectively.
As mentioned above the role of ArmRPs in nature is closely linked to their ability to bind extended peptides or unfolded parts of proteins. Importin-α, for example, recruits the nuclear localisation sequence (NLS) and binds it in an anti-parallel manner with the C-terminus of the peptide positioned towards the N-terminus of the armadillo protein

(see Figure 1.3 C). Crystal structures show that NLS lies at an angle of approximately 45° across the concave binding interface formed by the helices 3.

1.2 Protein Engineering and Consensus Design of Repeat Proteins

1.2.1 Overview of Protein Engineering

Protein engineering efforts aim at developing useful proteins with new or enhanced functions. This can be done based on the structural basis of existing scaffold proteins, which are then adapted and optimized for the required properties. Often these efforts try to make proteins more robust or aim at improving or altering substrate specificity and affinity e.g. of the catalytic centre of enzymes. Alternatively *de novo* design strategies have tried to build entirely new protein folds, which do not exist in nature [30].
Two general strategies are employed to achieve these goals – protein design and directed evolution. Often a combination of both strategies is used to obtain the desired results and can be complemented with input from computational methods.

1.2.2 Rational Protein Design

Protein design is a rational design approach drawing on existing knowledge to create proteins with new functions. It requires a detailed understanding of the target and can be a convenient and relatively inexpensive way to achieve a change in function, for example, by site-directed mutagenesis of a few residues. Naturally there is a certain limit to the amount of work which can be invested in a certain target [128]. Additionally the effects of sequence changes are often not easy to predict even when detailed structural information is available.
Another rational design approach which is commonly used as a basis for computational or experimental optimization is to establish consensus sequences based on amino acid sequences of similar or homologous proteins from different species to deconvolute conserved patterns determining the stability, fold and function of these proteins. This approach has been successfully used by the Regan and Plückthun groups to design consensus sequences of TPR repeat proteins [31] and Ankyrin repeat proteins respectively [32] (*vide infra*).
Also *de novo* design approaches fall under this category as they take advantage of the wealth of knowledge about protein folding, biophysical properties of amino acids available from databases. Efforts can range from protein redesign – recalculations of amino acid variations of a known structure while keeping the original fold – to the design of a fold and function unknown in nature. The protein Top7 was the first completely *de novo* designed protein fold [30]. *In silico* predictions of sequences adopting a certain fold are then validated experimentally and can be further improved by alternating between simulations and experiments.

The necessity to rely on the amino acid composition for rational protein design [33] has been overtaken by the availability of a vast number of high-resolution structures giving

access to detailed structural information. Furthermore, the development and decreasing cost of sophisticated computational methods such as improved protein design algorithms and explicit water simulation force fields in combination with the availability of large databases of protein folds, amino acid conformations, and other associated data have strengthened the field of rational protein design. A fast and accurate energy function, which can quickly distinguish between improved or suboptimal thermodynamic contributions, is a crucial requirement for the economical use of computational methods. Several novel enzyme functions have been established by the Baker and Hilvert groups using *de novo* and protein redesign approaches [34–36].

1.2.3 **Directed Evolution and Ribosome Display**

The second major strategy in protein engineering is directed evolution. This process includes cycles of random mutations and selection or screening, mimicking the natural process of evolution. The extent of randomization can be limited to a few sites and specific amino acid types or cover a large number of sites and amino acids. This approach can cover arrangements which are not easily conceivable by rational design and therefore often result in surprising but valuable results. Theoretically, no structural knowledge of the target is needed, however a limited understanding of the scaffold at hand can stream-line the process using a combined effort utilising directed evolution and rational design.
For example directed evolution has been successfully used to improve the affinity of antibodies [37] over several cycles of ribosome display maturation. Additionally, features of successful mutants can be further combined by shuffling DNA sequences, another evolutionary optimisation process borrowed from nature.

Both rational design and directed evolution strategies have profited from increased access to high-throughput technology. Particularly in the case of directed evolution experiments where a vast number of mutations and their combinations are to be exhaustively treated (typically 10^6-10^{12}) to cover the complete sequence scope. Large numbers of mutations are handled in the form of DNA libraries which need to be translated into proteins to be selected based on the target properties. One of the biggest challenges for successful selection or screening is to create assay conditions that encourage the desired protein activity. Only the correct biochemical environment during this process will lead the discovery of stable proteins with the desired target function.

Choosing the most suitable method of directed evolution for each target is crucial for success. The mantra “You get what you select for” has been repeatedly invoked to stress the importance of the initial experimental conditions (pers. com. A. Plückthun).

The armadillo repeat protein binder VG_328 investigated during this work (see Chapter 5) was established using ribosome display, an *in vitro* directed evolution method. This method is based on *in vitro* protein translation, leading to a physical complex of the genetic information, mRNA, and its translated product, the protein. The coupling of

genotype and phenotype is achieved by preventing the protein and mRNA from leaving the ribosome. The most prominent advantage of this method over other *in vivo* techniques such as phage display [38,39] or yeast display [40] is that it does not involve the otherwise necessary transformation of DNA into living cells – a process which is usually the bottle-neck regarding library size and therefore the potential to undertake exhaustive amino acid randomization. All methods involving a transformation step to bring the genetic information into cells face the difficulty of transfering the true size of the library into cells and, due to physical limits, often end up under-representing the original library's diversity - a library using a protein of 100 amino acids in size with 12 fully randomised amino acid positions theoretically encodes $3.8x10^{21}$ individual members – sample consisting of a single example of each library member would constitute ~50 g of protein! (based on average MW of 110 for each amino acid). The use of fully *in vitro* methods like ribosome display enables the selection of very large libraries without compromising diversity.

A further advantage of the ribosome display method is that apart from the initial mutations included in the original library additional random mutations can be introduced during *in vitro* amplification. This process continuously increases the scope of mutations with each cycle, enabling access to sequences which may not have been present during the initial library construction, and making this a truly Darwinian evolution process, while enriching proteins of the desired characteristics. As this added evolution step of mutations is carried out in vitro, the location, average frequency and residue type of mutations [41,42] can be controlled by well-established protocols depending on the desired effect. In addition libraries can be shuffled to recombine established mutations [43,44]. In contrast methods like phage display merely select from an existing pool of mutants provided by the original library.
Ribosome display has been successfully used to create proteins of higher affinity [37], higher stability [45] or other enhanced features. The principals of Ribosome display are summarized in Figure 1.4.

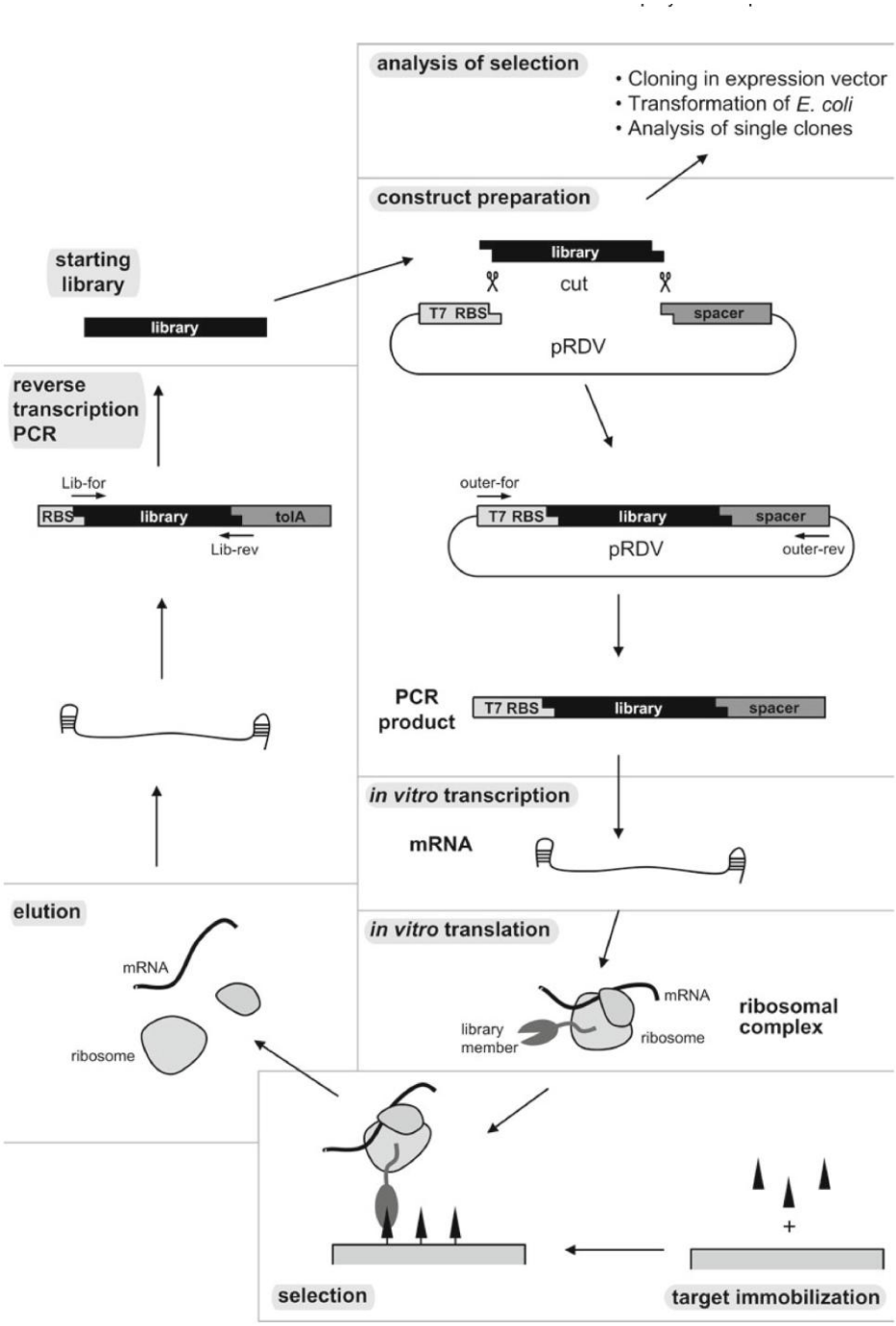


Figure 1.4 The principals of ribosome display: The target randomised DNA library is ligated into a specialised ribosome display vector, pRDV, containing a strong promoter, ribosome binding site and an in-frame spacer sequence. The library insert and the flanking regions are PCR amplified and transcribed *in vitro* into mRNA. This mRNA pool is translated *in vitro* using ribosome extracts. Translation of each mRNA stalls due to the presence of a stalling sequence contained in the mRNA thereby linking the freshly synthesized and folded protein to the mRNA via the ribosome. These complexes can now be subjected to affinity selection using immobilised targets. mRNA is extracted from the complexes retained by the target by dissociating the ribosomes. This mRNA pool is transcribed into DNA and

(Figure 1.4 cont.) amplified by PCR. During this amplification step further mutations can be introduced. The obtained pool of selected binder sequences can be used for the next round of affinity maturation. Several rounds of ribosome display can be carried out until a satisfactory enrichment is observed. The selected pool of DNA can then be cloned into expression vectors and used for *E. coli* transformation and small scale *in vivo* protein expression and screening of selected binders. As shown in "Ribosome display: A Perspective" [46].

1.2.4 The Design and Applications of the DARPin Scaffold

The availability of a large number of AR protein sequences has enabled the design of consensus-based scaffolds. These artificial proteins have been found to fold stably and display very high expression levels of soluble monomeric protein in *E.coli* [21].

The Plückthun group has undertaken an extensive design effort to create a highly stable designed consensus AR protein (DARPin) scaffold [32,47,48]. The caps, protecting the array of internal repeats from the solvent are, of enormous importance and were the subject of intense optimization efforts using computational approaches [48]. These studies have revealed positions responsible for scaffold integrity and protein interaction leading to an optimized design that provides randomisable surface residues to tailor the binding surface to a desired target without compromising scaffold stability. This scaffold has been successfully used to establish protein binders from randomized libraries against numerous targets and have found application as biologic therapies among other things [49,50].

Today these DARPins present a widely used scaffold for molecular recognition and are usually more thermodynamically stable than their natural counterparts and most globular proteins [21,51].

1.2.5 Scaffold Choice for a Modular Peptide Binding Platform

1.2.5.1 Overview

The identification and binding of peptides is a key element of intra- and extra-cellular signaling and transport pathways. The sequence-specific binding of peptides is of great interest for a wide range of academic and commercial applications. Proteomics, structural biology, medical diagnostics and biologics for therapy would benefit from fast access to sequence-specific binders of target peptide sequences.

Evolution has provided us with a plethora of peptide binding molecules, from immunoglobulins to small adaptor domains like SH3 domains, and even a number of peptide binders that utilise repeating structural units such as the WD40 domains, tetratricopeptide domains, and armadillo repeat proteins.

Currently, traditional monoclonal or recombinant antibodies [52,53] as well as alternative scaffolds like DARPins [32], Affibodies, Anticalins, and Adnectins and are widely used scaffolds to create binders of high affinity and specificity [54,55]. However, the bottle-neck of the current way of establishing binders is that each new binder has to undergo affinity

maturation, characterisation and testing for specificity for every new target, making the process time-consuming and expensive. Additionally, these scaffolds focus on binding three-dimensional epitopes like surfaces of folded proteins, rather than a specific peptide sequence.

Two important characteristics form the basis of our design approach: Firstly, in order to bind a peptide in a sequence-specific manner the peptide must be bound in extended conformation. Secondly, a modular building block-like system of pre-established units each recognizing a specific single, di- or tripeptide unit could circumvent the tedious process of establishing a new binder for each new target. A binder could then be simply constructed on demand from building blocks of a known specificity. A suitable scaffold must therefore fulfil these two criteria – extended binding mode and modularity – as well as display high binding affinities, high yields and robust behaviour *in vitro* and *in vivo*.

1.2.5.2 Evaluation of natural peptide binding proteins

Structural analysis of antibody-peptide complexes show that despite the undisputed advantage of high specificity and affinity, peptides are bound in a range of different conformations lacking a conserved binding mode [56–59]. Furthermore, their binding mode is not modular with peptide binding often being moderated by the variation of residues that lie on adjacent helices on either side of the binding groove, parallel to the axis of peptide binding (see Figure 1.5). Additionally, antibodies and their derivatives contain labile disulfide bonds rendering them unsuitable for intracellular application.

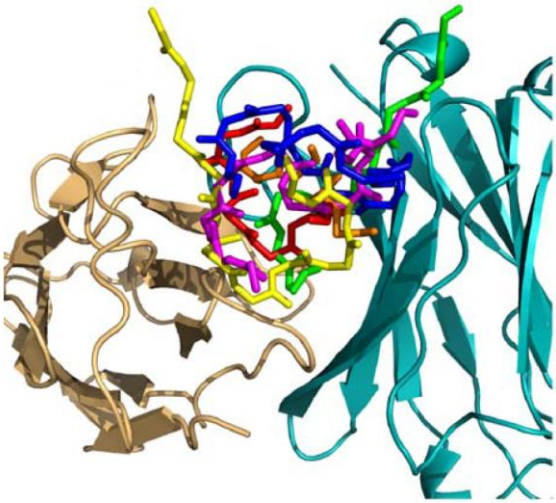


Figure 1.5 Peptide binding by an antibody, heavy and light chain are depicted in brown and light blue, the non conserved peptide conformations are shown as sticks. Figure as shown by F. Parmeggiani [60].

Small adaptor domains like SH3, SH2, PDZ and WW [61] bind peptides in a more conserved binding mode within protein families but often lack high affinity as their

origin in regulating cellular processes requires reversible interactions [62,63]. The bound sequences are very short and biased regarding the recognition of posttranslational modifications and certain amino acid types. The extension of the binding surface by linking several small adaptor domains via flexible linkers is not feasible due to the entropy penalty incurred through decreased scaffold flexibility upon binding of a peptide. These characteristics make it unlikely that their binding mode could be extended to bind longer peptides of an arbitrary sequence with sufficiently high affinity.

The major histocompatibility complex proteins MHC-I and MHC-II present peptide sequences of endogenous and exogenous origin to T-cells [64–66] and form an important part of the mammalian immune response. The peptide binding domain of MHC-I can only accommodate short peptides of 8-9 amino acids, while MHC-II has an open binding groove which can bind peptides of up to 20 amino acids in length. Unfortunately although they can bind a wide range of peptide sequences, the fact that MHC-I and MHC-II are heterodimeric, membrane bound complexes of low stability makes these proteins a very challenging scaffold for protein design [67].

Repeat proteins are interesting scaffolds especially when considering the desired modularity of the envisaged peptide-binding platform. The logic for utilising a repeat protein as a scaffold for generating engineered selective peptide binders is based on capitalising upon their modular nature to produce specific peptide binders which can be customised simply by adding, removing or shuffling repeats in the same manner as the target peptide may be changed by increasing its length or changing its sequence. As discussed in Section 1.1, repeat proteins have a large surface to volume ratio in comparison to globular proteins, which makes them optimal candidates or the development of a binding scaffold [2,5].
Several repeat proteins families possess an intrinsic ability to bind peptides on binding surfaces based on their repetitive nature. Four major peptide binding repeat protein families are the β-propeller proteins [68] (WD40 and Kelch domains), tetratricopeptide repeat proteins (TPR) [69], HEAT repeat proteins [70,71] and armadillo repeat proteins [72] (see Figure 1.6).

β-propeller proteins display a closed repeat structure and can bind proteins, peptides and DNA alike (Stirnimann et al., 2010). Peptides are usually bound across the upper surface of the channel formed by 4-10 repeats (“blades”, see Figure 1.6 A), each made of a four-stranded anti-parallel β-sheet, and are often post-translationally modified [73]. However, attempts to design stable β-propellor scaffolds have proven very challenging [74].

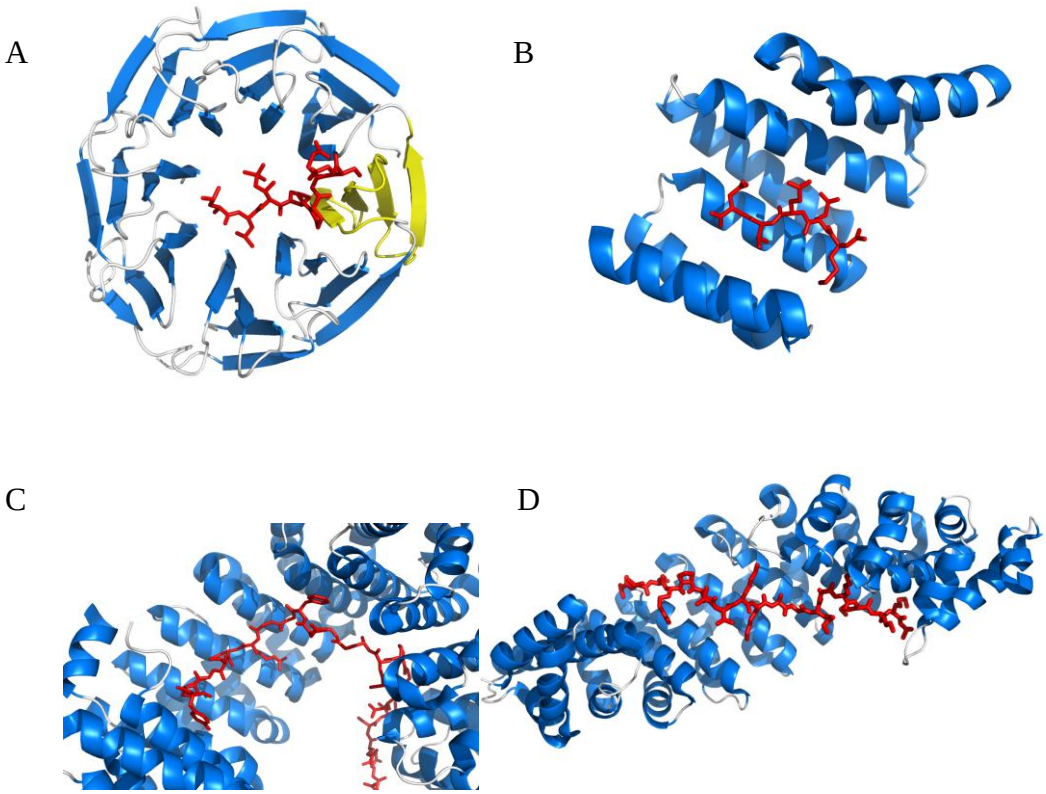


Figure 1.6 The peptide binding modes of four major repeat protein families. A: WD40 repeat protein (PBD ID 1NEX), B: TPR protein (PDB ID 1ELR), C: HEAT repeat protein (PDB ID 2H4M), D: Armadillo repeat protein (pDB ID 1EE5).

Tetratricopeptide repeat proteins are a well-studied repeat protein family [75] and have been used in several protein enigneering studies. The array of helix-turn-helix motifs forms a superhelical structure with a concave binding surface. Peptide ligands of varying sequence and secondary structure are bound by binding pockets along this surface (see Figure 1.6 B). Engineered modules based on tetratricopeptide repeats, which can bind short peptide motifs have been successfully established by the Regan group [76–78].

The HEAT repeat motif contains two helices (A and B) forming a helical hairpin, and is structurally similar to the armadillo repeat motif (see Figure 1.6 C). In HEAT repeat proteins helix A is bent, corresponding to helices 1 and 2 in armadillo repeat proteins [11]. In HEAT repeat proteins the array of all helices B forms the concave binding surface (helices 3 in armadillo repeat proteins). Even though the binding surfaces of the HEAT repeat proteins importin-β1 and –β2, for example, are highly conserved it is challenging to establish a HEAT repeat protein consensus sequence across subfamilies. This has limited engineering efforts to a small sequence subset [79].

Armadillo repeat proteins and their peptide binding abilities have been introduced in Section 1.1.3. Their right-handed super-helical structure includes a concave binding surface (see Figure 1.6 D), made up from the surface of the adjacent third helices of each repeat, similar to HEAT repeat proteins [71]. Even though identifying a common

consensus sequence across different subfamilies and species is not a straight-forward process, all armadillo repeat protein subfamilies share a common binding mode in which peptide sequences are bound in an extended conformation [80,81] (see Figure 1.7 A). These peptide sequences can be unfolded termini, loops of proteins or free peptides. Peptides are bound in an anti-parallel manner with the C-terminus of the ligand orientated towards the N–terminus of the armadillo repeat protein.
The nature of the binding groove varies slightly depending on the subfamily. In importin-α the binding surface contains a conserved ladder of asparagines (see Figure 1.7 B), which can bind the backbone of extended peptides orientating the ligand on the binding surface in a conserved fashion [29]. Across the binding surface each helix 3 contains a conserved asparagine residue at position 37 of the repeating unit. In the array of tandem repeats these aspargine residues create a sloping "ladder". The asparagine side chains can form strong bidentate hydrogen bonds to the backbone of a bound peptide leading to a conserved binding conformation. Additionally conserved tryptophan residues above the asparagine ladder (position 33) provide hydrophobic binding pockets for aliphatic side chains of a peptide ligand. In contrast in β-catenin this ladder contains not only asparagines, but also a histidine and a glutamine [11].

The key feature of this interaction, which forms the basis of our design efforts described in this thesis, is the conserved mode by which importin-α binds the backbone of NLS. Crystal structures of ArmRP-peptide complexes show that with the backbone fixed in this manner each repeat unit can bind two amino acid side chains. Interactions of the peptide side chains with the protein surface contribute further to the very high binding affinities of up to 10 nM [82] seen in natural armadillo-peptide interactions.
Because the tandem arrays of repeats in ArmRPs form a continuous domain, longer peptides can be bound by the modular binding surface by adding consecutive repeats. Up to three consecutive repeats have been found to bind peptides in this conserved manner in nature. Figure 1.7 illustrates this binding mode for importin-α binding the backbone of the NLS peptide with six hydrogen bonds. In nature, longer peptides are bound with a combination of the described binding mode and non-conserved interactions for example linking several binding sites as observed for the N- (major) and C-terminal (minor) binding sites in the importin-α : NLS complex.

In conclusion an inspection of the binding modes of these four important peptide binding repeat protein families (see Figure 1.6) shows that the β-propeller and tetratricopeptide domains have sub-optimal binding alignments of the peptide with respect to the repeats for full utilisation of the modular effect. The HEAT and the armadillo domain both show an extended peptide conformation where the peptide lies more or less perpendicular to the axis the binding face helix of each repeat and parallel to the axis of repeat stacking, enabling each repeat to selectively bind individual parts of the peptide. We chose armadillo repeat proteins over HEAT repeat proteins because we found them to be more amenable for the establishment of a consensus sequence.

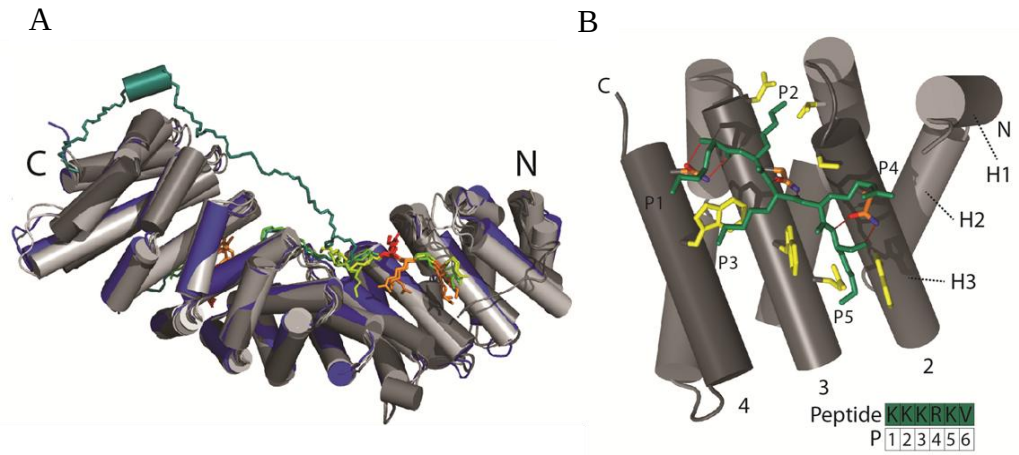


Figure 1.7 Details of the binding mode of importin-α as shown by Reichen *et al.* [83]. The α-helices of different ArmRPs (colored from grey to blue, with N- and C-termini labeled as N and C, respectively) are represented by cylinders. Peptides, bound in antiparallel orientation, are shown in sticks, colored from yellow to green. (A) Structural superposition of five members of the importin-α family, representing classes of peptides with different structural homology (PDB ID: 1EE5 [84], 1EE4 [84], 1Q1S [85], 1IAL [86], 3TJ3 [87]). Highly conserved peptide residues are highlighted by their side chain (colored from red to yellow). B: Detailed view of the major binding site of *yeast* importin-α (PDB ID 1BK6 [29]) in complex with the NLS peptide (green), making six backbone hydrogen bonds with the conserved Asn residues (orange). Residues of importin-α interacting with side chains of the peptide are shown in yellow.

1.2.5.3 Designing a Modular Peptide Binding Scaffold Based on Armadillo Repeat Proteins

As detailed in the section above armadillo repeat proteins are a naturally good choice as a scaffold for designing peptide binding proteins. The highly conserved nature of the binding mode is illustrated in the superimposed structures of importin-α and β-catenin binding various natural peptides and compared with an antibody, as shown by F. Parmeggiani [60] (see Figure 1.8)

ArmRPs from different families based on the same overall fold can exhibit specific binding affinities for oppositely charged target peptides. This indicates that binding specificity and scaffold fold are independent features in ArmRPs. The conserved modular binding mode provides the basis for sequence specific binding. To artificially propagate this conserved binding mode over more than what has been observed in nature (up to three repeats) the curvature of the super-helical binding surface must ensure strain-free binding of longer peptides. Curvature analyses of natural and designed ArmRPs indicate that the binding of longer peptides should be achievable [83,88].

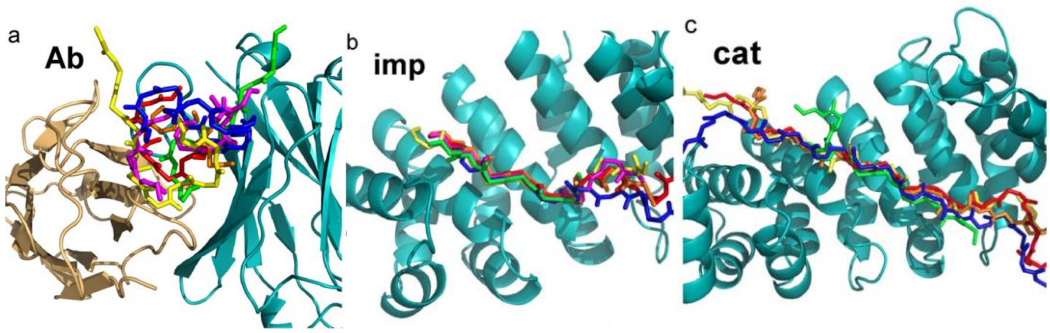


Figure 1.8 Comparison of peptide binding modes between antibodies (a) and the armadillo repeat proteins importin-α and β-catenin. Proteins are shown as cartoons, peptides are shown as sticks. Figure as shown by F. Parmeggiani [60]

In nature, the binding sites for the peptide side chains specifically accommodate different types of amino acids, however, a preference for charged amino acids has been observed. Importin-α binds the NLS sequence containing positively charged Arg and Lys residues [89], whereas β-catenin binds to negatively charged Asp and Glu residues. In a designed binder this natural preference would have to be extended to accommodate a broad range of amino acids for a truly universal peptide-binding platform. Furthermore consecutive Pro residues in a peptide lacking the backbone amide moiety would be less likely to bind in the conserved mode. A possible solution to this problem would be borrowing features from other binding domains like SH3 and WW-domains providing several parallel aromatic residues to enable peptide binding proline-rich peptides in a PPII-type helix conformation. (Zarrinpar et al 2003).

Theoretically, a range of individual repeats exhibiting affinity for specific dipeptide units could be established by screening and selecting in the context of longer peptides (see Figure 1.9). After successfully establishing a number of repeats of different specificity one could re-arrange several such building blocks into a new peptide binder of predetermined specificity using previously established repeats as stepping stones.

This would bypass the tedious and time-consuming process of affinity maturation for each new target and provide a valuable short-cut to the generation of specific peptide binders. Obviously, achieving this ambitious goal requires an interdisciplinary approach to design, characterise and fully assess sequence-specific ArmRP-based peptide binders.

Several milestones in creating the envisioned peptide binding platform have already been successfully established. These results have been recently reviewed in detail by Reichen *et al.* [83].

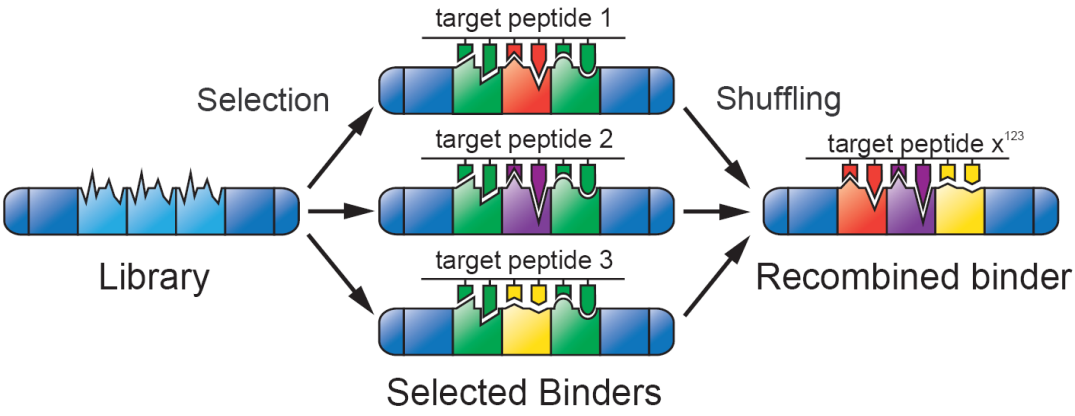


Figure 1.9 Schematic illustrating the strategic combination of library selection and repeat shuffling to establish repeat modules of defined affinity for dipeptide units. Figure adapted from Reichen *et al.* [83].

Table 1.2 Biophysical analysis of natural and designed armadillo repeat proteins. Table as shown by Reichen *et al.* [83].

Constructs[a]	Ref	Residues (repeats)[b]	pI[c]	MW_{calc} (kDa)[d]	Oligomeric state[e]	MW_{obs} (kDa)[f]	$MW_{obs/calc}$[g]	CD_{222} (MRE)[h]	T_m (°C)[i]	CD GdmCl (M)[j]
$Y_IC_4A_I$	1	253 (6)	5.1	26.9	Monomer	50.2[k]	1.86	-12,393[k]	n.d.	n.d.
$Y_IM_4A_I$	1, 2	253 (6)	5.1	27.1	Monomer	32.3	1.19	-19,255	71	3.5
$Y_I\overline{M}_4A_I$	2	253 (6)	4.5	27.1	Monomer	32.3	1.19	-19,162	76	3.7
$Y_{II}\overline{M}_4A_{II}$	2	252 (6)	4.5	26.9	Monomer	31.2	1.16	-20,401	86	4.4
$Y_{II}\overline{M}_3A_{II}$	2, 3	210 (5)	4.6	22.6	Monomer	27.5	1.22	-19,015	77	3.6
$Y_{III}\overline{M}_3A_{II}$	3	210 (5)	4.8	22.5	Monomer	27.4[k]	1.22	-20,259[k]	81	3.8
$Y_I\overline{M}_5A_I$	4	295 (7)	4.4	31.5	Monomer	38.6	1.23	-20,435	80	4.2
VG_328[l]	4	295 (7)	4.4	31.7	Monomer	39.9	1.26	-20,199	74	3.3
Importin-α[m]	1	435 (10)	5.5	48.2	Monomer	43.0[k]	0.9	-14,646[k]	43	n.d.
β-catenin[n]	1	528 (12)	8.7	57.6	Monomer	52.8[k]	0.9	-17,207[k]	58	n.d.

[a] Capping repeats ($Y_I/Y_{II}/Y_{III}$ and A_I/A_{II}) and internal repeats (C, M, $\overline{M}$) are given in Figure 4.

[b] The number of residues includes the $MRGSH_6GS$ tag; the number of repeats includes capping repeats.

[c] Isoelectric point (pI).

[d] Molecular weight calculated from the sequence; masses were confirmed by mass spectrometry.

[e] Oligomeric state as indicated by multi-angle static light scattering.

[f] Observed molecular weight as determined by size exclusion chromatography.

[g] Ratio between observed (size exclusion chromatography) and calculated molecular weight ($MW_{obs/calc.}$)

[h] Mean residue ellipticity at 222 nm expressed as deg cm^2/dmol.

[i] Transition midpoint (T_m) observed in thermal denaturation measured by CD.

[j] Midpoint of transition in GdmCl-induced denaturation, measured by CD.

[k] Normalized to value from Alfarano et al., 2012.

[l] Binder VG_328 is derived from a $Y_I\overline{M}L_3\overline{M}A_I$ library, where L is a randomized library module.

[m] Armadillo domain of human importin-α1.

[n] Armadillo domain of mouse β-catenin.

[1] Parmeggiani et al., 2008

[2] Alfarano et al., 2012

[3] Madhurantakam et al., 2012

[4] Varadamsetty et al., 2012

As the amount of sequence and structural information increased, the armadillo repeat protein family was divided into subfamilies and crucial residues in the consensus sequence were more clearly defined in 2001 by Andrade *et al.*[11]. Parmeggiani *et al.* established the first artificial armadillo repeat proteins based on consensus sequences of homologous proteins of the importin-α (Type I) and β-catenin (Type T) subfamilies and a combination (Type C) of the two [90](see Figure 1.10). Establishing a consensus sequence was an especially challenging process due to the low sequence identity between the different subfamilies. Capping repeats were either taken from natural sequences (Y-cap = N-cap from natural **y**east importin-α) or artificially designed (A-cap = C-cap, **a**rtificial).

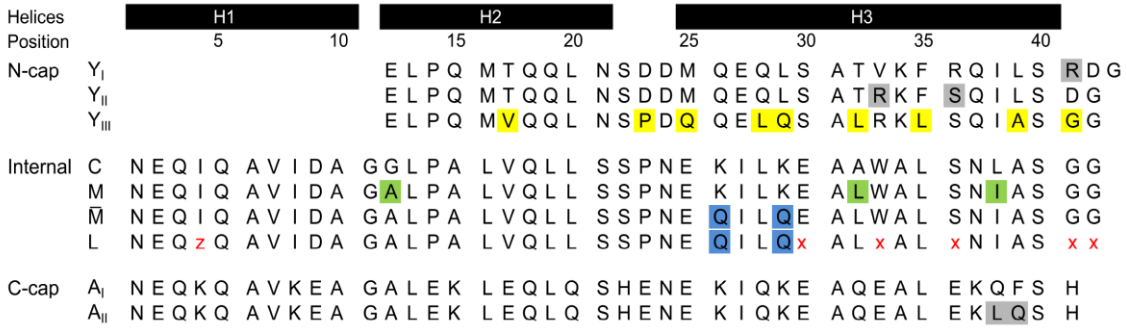


Figure 1.10 Sequence alignment of the N-caps (Y_I, Y_{II} and Y_{III}), internal repeats (C, M, M̄ and L) and C-caps (A_I and A_{II}). Mutations introduced in the internal repeats to improve packing of the hydrophobic core and decrease charge repulsion on the surface are colored green and blue, respectively. Residues mutated or deleted in the terminal caps upon MDS analysis are colored grey. N-terminal cap mutations introduced to prevent domain swapping are colored yellow. The nomenclature for designed Armadillo Repeat proteins assigns two α-helices (shown in black rectangles) for the N-terminal cap, and three for the internal repeats and C-cap, respectively. Experimentally tested proteins consisted always of an N-cap, several internal repeats and a C-cap (e.g. $Y_IM_4A_I$). Library module L is based on the internal repeat M̄ and contains in total six randomized residues (z: E,H,K,I,Q,T or R and x: all 20 amino acids except P,C and G). Figure as shown by Reichen *et al.* [83].

Proteins were expressed in a $Y_zC_xA_z$ format, z denotes the design generation of the capping repeats (roman numerals I-III) and x represents the number of internal repeats used. Remarkably, these designed proteins were solubly expressed at high levels in *E. coli*, however, they were found to have dimeric or molten globule-like characteristics Table 1.2.

The initially established C-type consensus sequence could be improved using computational modeling of $Y_IC_4A_I$. By using molecular dynamics simulations including cyclic heating and energy minimization improved packing of the hydrophobic core of the internal C-type repeats could be achieved. The newly established repeat type was

named M-type for **m**olecular dynamics. $Y_I\mathbf{M}_4A_I$ proved to be soluble, stable and monomeric and exhibited cooperative unfolding behaviour [90] (see Table 1.2).

A persisting feature of proteins containing M-type repeats could be uncovered by Alfarano et al. by tracking the quality of [^{15}N,^{1}H]-HSQC spectra at different pH values. Broad signals were observed below pH 10, indicating insufficient side chain packing. Using a range of biophysical techniques, molecular dynamics simulations and NMR spectroscopy, two lysine residues in each repeat could be identified in closely spaced positions leading to unfavourable repulsion. Consequently these lysines (K26 and K29) were mutated to glutamines, which lead to improved characteristics independent of the pH. The modified internal repeat type was named $\bar{M}$ [91] (see Chapter 3).

As previously observed for Ankyrin repeat proteins [47,92], apart from the requirement to design stable internal repeats, the design of optimal capping repeats proved to be of the utmost importance for protein stability. Using a similar interdisciplinary approach the capping repeats could be successfully re-designed leading to second generation caps (II). Proteins of the format $Y_{II}\bar{M}_3A_{II}$ and $Y_{II}\bar{M}_4A_{II}$ were considerably more stable than those with caps of the first generation (see Table 1.2) [91]. The effects of the individual mutations introduced in this design cycle are discussed in detail in Chapter 3 and 5. In agreement with previous findings for other repeat proteins, protein stability of these constructs increased with increasing repeat number [90,93–96].

Madhurantakam *et al.* successfully crystallized proteins containing the improved capping repeats and internal repeats ($Y_{II}\bar{M}_3A_{II}$ and $Y_{II}\bar{M}_4A_{II}$) providing the first structural information of designed armadillo repeat proteins [88]. These crystal structures showed the typical solenoid fold of the armadillo domain, however, under crystallizing conditions the N-cap (Y_{II}) engaged in a domain swap forming a dimer with the N-cap of a neigbouring molecule and did not fold as intended by the design. Closer analysis showed that the consensus design of the internal repeats was lacking a helix breaker, which is usually present in the first repeat of natural armadillo repeat proteins. To enforce proper folding of the N-cap, the third generation Y_{III} was established by likening the sequence of the N-cap to that of the internal repeats. The resulting $Y_{III}\bar{M}_3A_{II}$ protein was crystallized as a monomer with a properly folded N-cap and gained further thermal and chemical stability [88] (see Table 1.2).

The designed armadillo repeat proteins discussed so far did not exhibit high affinities for the NLS sequence (KKKRKV). This was surprising as the minor binding site of importin-α (PDB ID 1BK6) is very similar to the structure found for the internal repeats in $Y_{III}\bar{M}_4A_{III}$ and most important residues involved in binding NLS are present in the consensus sequence. Closer analysis revealed that the conserved binding pocket P1' is missing in the consensus. Based on this observation rational design to introduce such a binding pocket can be undertaken.

In parallel Varadamsetty et al. used a directed evolution approach and established randomized libraries based on the consensus design described above. Initial libraries were set-up in a $Y_I\bar{M}_3A_I$ format with first generation caps. Randomized positions were chosen based on crystal structures of natural armadillo repeat proteins and their ligands. Six positions in total were randomized, five in helix 3 and one in helix 1 of each repeat (see Figure 1.10). Even though Gly, Pro and Cys were excluded for randomized positions and only a limited subset of amino acids was allowed for helix 1, randomization still lead to lower protein stabilities in library members. To counteract this, two unrandomized repeats were introduced flanking the randomized repeats to provide a more stable scaffold, which improved the biophysical properties of library members [95].

Ribosome display of the stabilized library format was carried out with the neurotensin peptide (NT, QLYENKPRRRPYIL) as target [97]. After four rounds the selected pool of proteins was screened for NT affinity using ELISA assays. Two peptide binders, VG_328 and VG_306, for NT were identified, which show different affinity for NT and only differ in one residue on the proposed binding surface. VG_328, the better binder of the two was analyzed further and showed comparable biophysical characteristics to unrandomized consensus proteins (see Table 1.2) and good specificity for NT in ELSIA assays. Binding could only be competed with NT and none of the other reference peptides tested. Additionally an alanine-scan of NT was performed and revealed that mainly four residues (P7, R8, R9 and Y11) of NT contribute to binding. Surface plasmon resonance measurements determined a moderate dissociation constant of 7 µM for NT [95]. The interaction of NT with VG_328, was investigated in more detail as part of this work (see Chapter 5).Folding of Ankyrin Repeat Proteins

1.2.6 **Ankyrin Repeat Protein Folding**

A detailed understanding of protein folding and protein-ligand interactions is vital for successful protein design. The process of how a peptide chain folds into its correct three-dimensional structure is of enormous interest to close the gap in our understanding of the informational flow from DNA to functional proteins. Understanding the underlying mechanisms of protein folding would have a transforming effect on the fields of protein *de novo* design and structure prediction algorithms, as well as help to address diseases caused by misfolded proteins.
Repeat proteins in general, and AR proteins in particular, are an interesting subject upon which to study protein folding. Due to their modular architecture repeat protein folds are governed by short-range interactions setting them apart from globular proteins, which are mainly stabilized by interactions between residues that are far apart in the sequence. Their low contact order and modularity represents an intriguing background against which to study the mechanisms of protein folding and protein stability.

Numerous stability and folding studies of natural and designed AR proteins have been conducted. Biophysical, crystallographic, molecular dynamics and NMR data have helped to understand the topology of AR proteins and how they mediate protein-protein

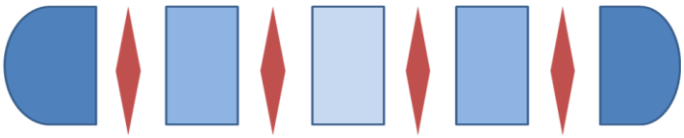
interactions. Furthermore, they provided insight into protein stability and folding pathways with many studies suggesting that this highly stable scaffold does not generally follow a two-state folding transition. For example, myotropin has been identified as a two-state folder following parallel folding pathways with either the N- or the C-terminal cap functioning as nucleation site [98,99]. However, stable intermediates have been observed for several AR proteins of less than optimal stability such as tumour suppressor p16 [100,101] showing that AR protein folding is not always a simple cooperative process. Typical intermediates contain unfolded capping repeats while the internal repeats are still folded. A general drawback of studying folding in natural repeat proteins is that the obtained conclusions can rarely be generalised due to inter-repeat sequence differences.
In conclusion, many thermodynamic equilibrium studies of AR proteins found simple two-state behaviour, whereas kinetic studies of AR proteins could detect intermediate states showing that the AR protein folding mechanism is more complex. This was confirmed in the context of consensus-designed DARPins by Wetzel *et al., vide infra*.

The consensus sequence of the AR has been robustly determined by several groups [19,21,32,51,94]. A consensus DARPin scaffold established by the Plückthun group [32] with identical repeats was used to study folding of DARPins in a uniform environment [96]. The same scaffold was used in this work for site-specific stability and folding studies of DARPins described in Chapter 2.
In their previous studies Wetzel *et al*. found that the stability increases with the number of internal repeats making proteins with more than three internal repeats more resistant against heat or denaturant induced unfolding. Analysing the folding and unfolding kinetics of protein members of up to 3 repeats of this consensus DARPin scaffold, they found folding to be monophasic with an almost identical rate for proteins of different length. In contrast, the unfolding rate decreased by a factor of 10^4 for each added repeat. They concluded that for folding the same transition state (e.g. the folding of one repeat unit) must be crossed regardless of protein length. The unfolding process, however, requires the unfolding of all individual folded repeat units, which explains the dependence of unfolding rates on repeat number. This is reflected in the unusually high thermodynamic stability of these DARPins, which require boiling in 5 M guanidine hydrochloride to unfold members of four internal repeats and higher.
Accordingly the number of unfolding phases increases with repeat number. An Ising-model (see Section 1.3.2) was found to describe both the thermodynamic and kinetic data very well. In addition they concluded that the thermodynamic stability of AR proteins and, potentially, repeat proteins in general is determined by their unfolding rate.
Following up on these results we investigated the stability and folding behaviour of this scaffold further using NMR spectroscopy (see Chapter 2 page 32). The sequence background of identical repeats used for our study can be seen as a generalised example for the study of AR protein folding and enables the investigation of folding as a function of repeat number.

1.2.7 **The Ising Model in Repeat Protein Folding**

As an alternative to classical cooperative folding models, one-dimensional Ising models have been successfully used to explain the unfolding behaviour of repeat proteins at equilibrium. In these models each repeat is treated as an independent folding unit. This can be useful to explain folding intermediates where some repeats are folded and others unfolded.

Originally, Ising models were used in statistical mechanics to describe the effects of magnetization. Ising-like models are used to describe thermodynamics in ferro-magnetic materials. These models were first utilized to describe folding behaviour of the natural ankyrin domain of Notch [102] and later for DARPins [12] and tetratricopeptide repeat proteins [93]. Ising models helped to deconvolute contributions towards overall stability from inter-repeat and intra-repeat stability, which allowed the optimization of these parameters individually. Using these models, the overall cooperativity of tetratricopeptide repeat proteins could be improved by stabilizing the interfaces between repeats by Main *et al.* [103], and Regan *et al.* used Ising models to predict mutations which improved intra-repeat stability in tetratricopeptide RPs without changing inter-repeat stability [104].



$\Delta G_{0N}' + J + . \Delta G_0 + J + \Delta G_0 + J + \Delta G_0 + J + \Delta G_{0C}'$

Figure 1.11 Schematic contributions to the free energy of unfolding according to the Ising model of protein folding as used by Wetzel *et al.* [96]. Blue boxes represent the intra-repeat contributions ΔG_0, red diamonds represent the interrepeat contributions J.

In protein folding, Ising-like models like the one used by Wetzel *et al.* [12] treat each repeat as an independent two-state folding unit (see Figure 1.9). The free energy ΔG_0 of unfolding of these units depends on the denaturant concentration in a linear manner. Additionally, the interaction of neighbouring repeats contributes a stabilizing potential, J. This potential depends on both repeats being folded, is independent of the denaturant concentration and of very high significance to DARPin stability.

Separate free energy terms, $\Delta G_0'$ are used to describe capping repeats which differ from the internal repeats in sequence and stability. The total free energy of the completely folded state of a three internal repeat consensus protein is defined by the sum of the free energies of the individual repeats and the caps, and the interaction potentials between all folded units. An intermediate state with an unfolded C-cap would lose the contributions of $\Delta G_{0C}'$ and the interaction potential J to the adjacent internal repeat.

A more detailed description of the model used in this study can be found in Wetzel *et al.* 2008 [12] and in Chapter 2 of this thesis.

1.3 NMR Spectroscopy of Proteins

1.3.1 Introduction

NMR spectroscopy is a powerful tool for the investigation of interactions, dynamics, folding and structure of proteins [105,106]. Its strength lies especially in the ability to characterise the dynamics of systems or interactions which are too flexible or transient to be detected by "snap-shot" techniques like X-ray crystallography. This is of particular interest for the analysis of weaker protein-protein or protein-ligand interactions.

For protein NMR, the NMR active nuclei ^{1}H, ^{13}C and ^{15}N are of particular importance. In a nut-shell, NMR spectroscopy determines the resonance frequencies of nuclei by exposing them to a static magnetic field and exciting them with electromagnetic pulses. Theoretically nuclei of the same type, ^{1}H, ^{13}C or ^{15}N have the same resonance frequency, however, in reality they vary depending on their chemical environment. These different resonance frequencies can be expressed as normalized chemical shifts (δ) in ppm in relation to a reference compound. Depending on the chemical nature of a compound the resonances from several types of NMR-active nuclei can be correlated leading to multidimensional spectra, for example a 2-dimensional [^{15}N,^{1}H]-HSQC.

Because the relative natural abundance of ^{15}N and ^{13}C isotopes is low (0.37 % and 1.1 %), multidimensional NMR experiments require that proteins are expressed from cultures grown using isotopically enriched media, which results in high levels of isotope incorporation into the expressed protein.

1.3.2 Heteronuclear Single Quantum Coherence - The HSQC

One of the fastest and most useful NMR experiments for the analysis of proteins is the [^{15}N,^{1}H]-HSQC (Heteronuclear single quantum coherence [107]. This experiment correlates resonances of moieties containing covalently coupled ^{15}N and ^{1}H nuclei. This conveniently gives rise to one signal for each non-proline peptide bond of a protein's backbone as well as for side chains containing amide groups like Arg and Asn, and Trp indoles. The HSQC of a protein shows a characteristic pattern of chemical shifts, identifying the protein's particular state like a finger-print. This finger-print provides valuable information on the protein's folding state. A well-dispersed pattern of signals above 8.5 ppm in the ^{1}H-dimension indicates a well-folded protein, whereas a narrow dispersal of signals (8-8.5 ppm for amide protons) can point out unstructured protein, which is freely sampling a large number of conformations or molten-globular states. The [^{15}N,^{1}H]-HSQC can also quickly provide feed-back reflecting effects due to changes to the protein such as single point mutations, ligand binding or changes in its oligomeric state.

1.3.3 **Resonance Assignment of Proteins**

The process of identifying which resonance signal arises from which nucleus in a protein is referred to as resonance assignment. Selective assignments of different parts of a protein can be obtained, with the backbone assignment being more easily accessible in comparison to the assignment of side chains.

1.3.3.1 Backbone Assignment

Traditionally the backbone assignment of proteins covers the proton and nitrogen nuclei of the amide moiety (NH), the carbon nuclei of the carbonyl moiety (CO) and the Cα (CA). Furthermore, the resonance of the Cβ nucleus (CB), which provides additional information on the amino acid type, is often obtained.

These nuclei can be assigned using a set of 3-dimensional triple-resonance spectra, which correlate the various ^{13}C-nuclei of each residue to the amide resonances of the [^{15}N,^{1}H]-HSQC. Using an experiment reporting on resonances of both the current residue i and the previous residue i-1 and a complementary experiment only reporting on resonances of the preceding residue i-1 (e.g. the HNCA and the HN(CO)CA), contiguous stretches of amino acid residues can be linked by matching the specific resonance frequencies of each residue to its preceding neighbour. These stretches can later be mapped onto the sequence of the target protein. Unfortunately, proline residues, which have no amino proton in the peptide bond, cause breaks in the backbone assignment. In most non-repeating proteins these gaps are easily overcome by means of exclusion based on stretches of unique sequence. When the same amino acid sequence, bounded at each end by a proline, occurs multiple times, as can be the case in repeat proteins, whole fragments can have multiple assignment solutions. These ambiguities require the use of additional spectra, particularly utilising the NOE effect, to bridge these gaps through space.

The standard set of experiments used for the backbone assignment and the correlations, which are obtained from them, are illustrated in Figure 1.10. A descriptive list of NMR experiments used in a particular project is given in the Materials and Methods Sections, and Supplementary Materials of the respective chapters.

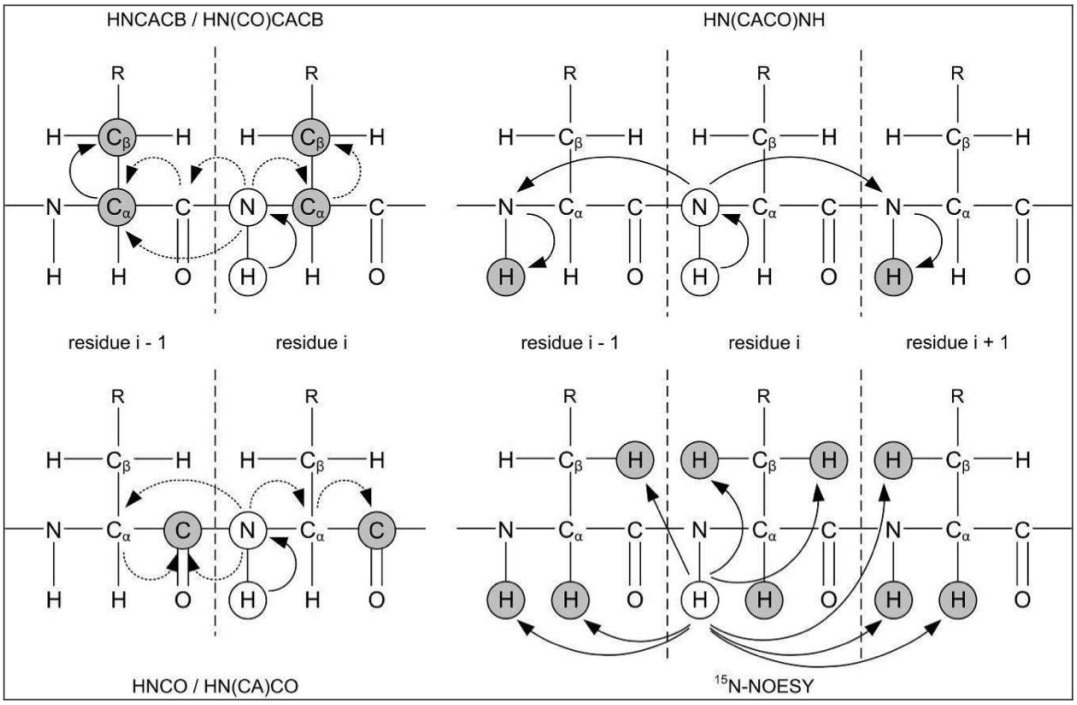


Figure 1.12 Illustrative summary of the main 3D NMR data sets utilised for the resonance assignment of backbone nuclei. The HNCACB [124] and HN(CO)CACB [108,124] set correlates the reference N-H (i) resonance to the Cα and Cβ carbon nuclei of residues i and i-1; the HNCO/HN(CA)CO [109,126,127] experiments do the same for the NH_i to the CO_i and CO_{i-1} resonances. The HN(CACO)NH [110] correlates the NH_i resonance to the NH_{i-1} and NH_{i+1} resonances, and the ^{15}N-NOESY utilises the NOE effect to generate distance restraints between NH protons and other protons through space. Figure adapted from Bumbak, 2012 [111].

1.3.3.2 Side Chain Assignment

Interactions in peptide-protein binding are often predominantly mediated by side-chain hydrophobic and charge-charge interactions, as well as by hydrogen bonds. In addition to binding information, the assignment of side chain-resonances, particularly the side-chain proton resonances, is required for the calculation of protein structures determined by distance restraints generated from ^{13}C- and ^{15}N-NOESY experiments.

To allow access to this information through NMR experiments the detailed assignment of side chain nuclei is necessary. The side chain assignment builds upon the backbone assignment linking the ^{1}H and ^{13}C resonances of side chain nuclei to the Cα and Cβ, and to the NH of their respective residues. The NMR experiments used to assign side-chain nuclei can be grouped into two different types: [^{13}C,^{1}H]-HSQC and [^{15}N,^{1}H]-HSQC based experiments. The (H)CCH-TOCSY and the ^{13}C-edited 3-dimensional NOESY are [^{13}C,^{1}H]-HSQC based, whereas the (H)CC(CO)NH and the H(CCCO)NH are based on the [^{15}N,^{1}H]-HSQC. Experiments based on the [^{13}C,^{1}H]-HSQC are traditionally used for structure determination, whereas experiments based on the [^{15}N,^{1}H]-HSQC can utilise previously established backbone assignments and provide convenient and fast access to

selected side-chain assignments but are of much decreased sensitivity and hence are limited to smaller proteins. These experiments allow the assignment of most side chain types and can be further supplemented by experiments reporting on chemical shifts of aromatic side chains. As above, the details of these experiments are included in the respective chapters.

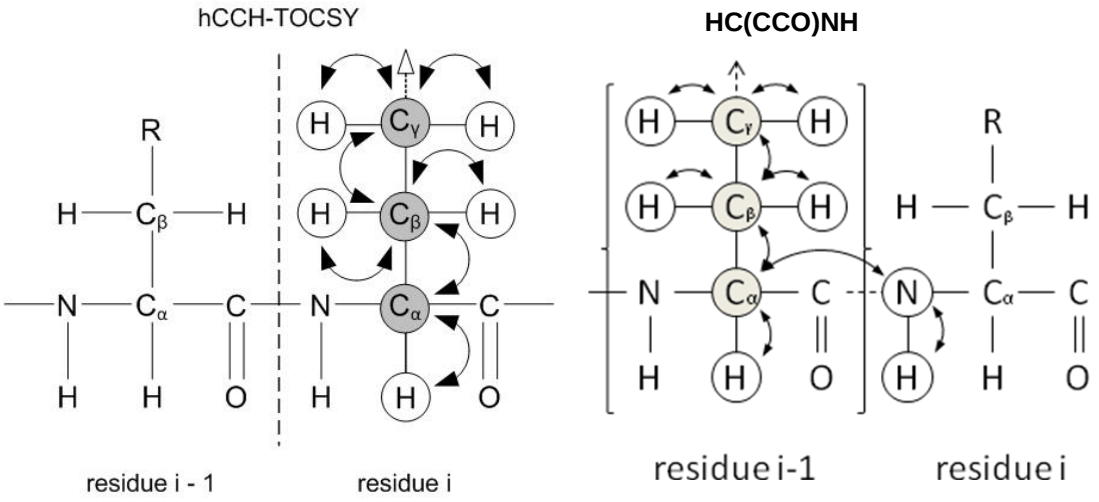


Figure 1.13 Correlations from 3D NMR experiments used to determine side chain resonance assignments. The (H)CCH-TOCSY shows correlations between the protons and carbons of the same side chain; B, The HC(CCO)NH shows correlations from the NHi to the i-1 sidechain H or C nuclei depending whether the experiment is conducted to detect protons or carbons. (H)CCH-TOCSY figure adapted from Bumbak, 2012 [111].

1.3.4 NMR of Repeat Proteins

An obvious advantage of multi-dimensional NMR spectroscopy is that resonances can be resolved along more than one axis, which reduces signal overlap in large molecules like proteins. The one-dimensional ^{1}H-spectrum of a folded protein is crowded, and the assignment of any peaks to any particular nucleus, impossible. Resolving the amide ^{1}H-resonances along a second dimension, ^{15}N, often allows almost complete deconvolution of the ^{1}H-resonances as each resonance is now clearly defined by two unique chemical shifts, $\delta\ ^{1}$H and $\delta\ ^{15}$N. In well folded, non-repeating proteins of up to ~30 kDa the combination of the amino and proton chemical shifts, as a result of the different electronic environment for each residue, often prove sufficient to avoid overlap of the resulting [^{15}N,^{1}H]-HSQC. However, in repeat proteins, where a sequence and the resulting structural motif is repeated several times, the aspect of a unique chemical environment is lost, resulting in highly similar to identical electronic environments for the nuclei at the same position in the different repeats. Sometimes the slight differences can be resolved using high-field spectroscopy, but often, especially in the case of engineered consensus repeats, where all repeats are of identical sequence, the resulting spectra are highly degenerate. Consequently the assignment of repeat protein spectra can

be a very challenging process. The traditional assignment approach described above can often only provide incomplete or ambiguous results.
A number of techniques have been developed to overcome this technical barrier.
Arguably, the most elegant of these techniques is the selective isotopic labelling of different repeats, so as to visualise only part of the whole protein molecule during any series of NMR experiments. This can be achieved using both biological and chemical methods, wherein the protein is synthesised in fragments, with only one part being isotopically labelled. These obtained fragments are later recombined forming a covalent link to create the complete protein, which is now labelled in one part but not in the other [112–116].
Alternatively,PRE-tags (see Section 1.4.6.2) can also be used to attenuate resonances from one part of a molecule in a distance dependant manner.

1.3.5 **Protein Stability Determination by NMR Spectroscopy**

1.3.5.1 Proton Exchange Experiments

Proteins are not static entities but are subject to internal movements and sit "breathing" at and around equilibrium in their folded state while small and very localized unfolding events can occur. This behaviour, the inherent flexibility and dynamics of a folded protein in solution, can be probed by NMR. A suitable measure for this purpose is the exchange rate of labile protons, for example amide protons with the aqueous medium. This measurement is undertaken by transferring the protein of interest, in fully protonated form, into a deuterated (D_2O) buffer and recording a time series of [^{15}N,^{1}H]-HSQC experiments, effectively tracking the loss of signals as protons are exchanged for deuterons from the solvent. Over time, all exchangeable protons will be exchanged for deuterons of the bulk solvent. Resonances arising from amide groups which are solvent exposed or situated in flexible loops and rapidly exchanging with the bulk solvent are often lost very quickly, whereas signals of residues buried in the hydrophobic core of the protein can decay considerably slower. Thereby a picture of the residue-specific stability of various parts of the protein can be established. This method was used to analyze the unusually high stability of DARPins described in Chapter 2.

1.3.5.2 The Heteronuclear NOE

The Nuclear Overhauser Effect (NOE) is a phenomenon based on the transfer of nuclear magnetisation via cross-relaxation from one population of nuclei to another [117–119]. The most useful feature of this effect is that it occurs through space, in contrast to spin-spin coupling, which requires nuclei to be linked by covalent bonds. Because this effect occurs in a distance dependant manner and is usually restricted to a maximum range of ~6 Å, it can be used to assess the spatial relationships between intra- or intermolecular non-bonded nuclei to establish the three-dimensional structure of a molecule [120]. This proved to be a major step in the advancement of protein structure determination by NMR, an approach pioneered by Nobel laureate Kurt Wüthrich.

Apart from its impact on structure determination, the NOE effect has also been exploited to investigate the flexibility of a protein's backbone by measuring the cross-relaxation rates of the backbone amide moieties in form of the heteronuclear $^{15}N\{^{1}H\}$-NOE. This simple 2-dimensional experiment based on a single refocused reverse INEPT sequence can provide valuable input for protein engineering efforts aiming at stabilizing a protein fold. Usually two different spectra, one with and one without the h-NOE contribution are recorded in an interleaved manner, which can be processed separately and used to determine a ratio of signal intensities obtained with and without proton saturation. The backbone ^{1}H-^{15}N heteronuclear NOE reflects the motion of individual N-H bond vectors in relation to the general motion of the whole molecule in solution. Faster motions of individual amide moieties result in lower NOE intensities, and are often observed for residues located in flexible loops and at the N- and C-terminus of a protein.

1.3.5.3 Chemical Denaturation Induced Chemical Shift Perturbation

The chemical shift of any resonance in an NMR spectrum is susceptible to the chemical environment of the nucleus from which the resonance arises. Changes in the electronic environment through changes in the protein structure or the presence or absence of a ligand can have a noticeable effect on the chemical shift of affected moieties. The perturbation of chemical shifts can be tracked in a concentration-dependant manner when a protein is subjected to increasing concentrations of chemical denaturant. For example, a series of $[^{15}N,^{1}H]$-HSQC experiments acquired at different denaturant concentrations can yield valuable information as to the stability and folded state of proteins under varying conditions. These experiments are especially useful to deconvolute the loss of tertiary structure and secondary structure during unfolding, as other biophysical methods such as CD-spectroscopy can only detect the presence or absence of secondary structure (see Chapter 2).

1.3.6 **Probing Weak Protein Ligand Interactions**

1.3.6.1 Ligand Induced Chemical Shift Perturbation

As mentioned in section 1.4.5.3 chemical shifts are susceptible to local environmental changes. These perturbations can also occur as a result of ligand binding, and can not only be informative of the nature and location of interactions between the protein and a ligand, but can also be used to determine binding equilibria of such interactions, as the degree of chemical shift change is relative to the degree to which the ligand is bound. Ligand binding can also cause secondary effects in the structure of a protein, for example by allowing a previously unstructured part of the protein to assume a defined structure. All these changes are mirrored by subsequent changes in the chemical shifts of the nuclei resonances. This is discussed in more detail in Chapter 5.

1.3.6.2 Paramagnetic relaxation enhancement PRE

Paramagnetic relaxation enhancement effects are widely used to obtain distance restraints in structure determination or, as done in this work, to confirm assignments of identical repeat units in a protein (see Chapter 2), or to identify the binding location of a peptide on a protein (see Chapter 5).

The use of paramagnetic relaxation enhancement to study protein-ligand interactions has been well described [121] and is a powerful method to characterise peptide interactions. The presence of unpaired electrons in an environment, such as those found in some metal ions and, for example, in nitroxide radicals, can lead to enhanced relaxation of NMR active nuclei. This phenomenon is referred to as paramagnetic relaxation enhancement (PRE) and effectively broadens signals by increasing relaxation rates of nuclei close to a paramagnetic moiety. This effect occurs through the dipolar interaction of an unpaired electron and a nuclear spin [122] in a distance dependant manner and can be described by the Solomon-Bløмbergen equation. A modified version of the equation more frequently used in biomolecular NMR was described by Battiste *et al.* [122].

PRE effects can be observed due to undesired contaminants like high levels of dissolved oxygen or metal ions in a sample,which can lead to broad signals and poor spectra quality. However, when introduced in a directed manner for example by incorporating paramagnetic metal ions into metal-binding proteins, valuable structural information can be obtained [123]. An alternative approach is site-directed spin-labelling, in which so-called spin-labels or PRE-tags are covalently attached to a protein or ligand to gather distance related information. A selection of PRE-tags commonly used for proteins is given in Figure 1.11

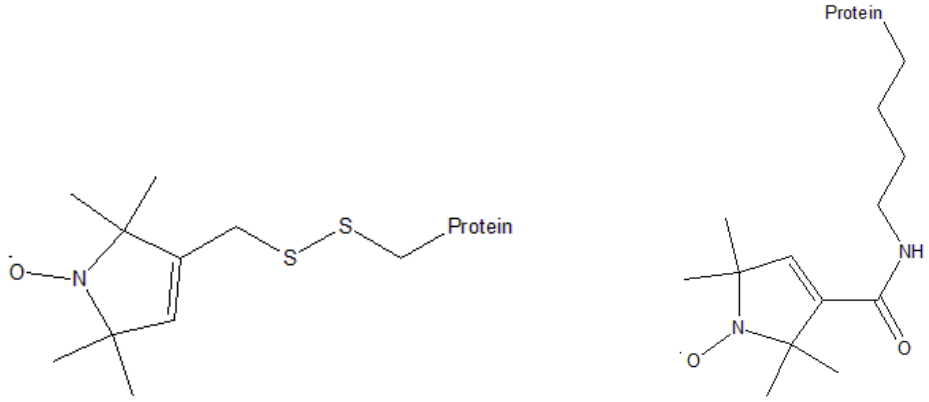


Figure 1.14 Chemical structures of the spin labels MTSL (left), for the labelling of cysteine residues and NHS (right), for amine e.g. lysine specific labelling.

As opposed to NOE restraints, which are limited to a maximum range of ~6 Å (see Section 1.4.5.2) PREs can provide valuable long-range distance restraints of up to 25-35 Å.

This is especially useful for large proteins, where NOEs can be difficult to obtain and assign, as PREs do not necessarily require side-chain assignments. In addition, input

from PREs can be used to detect weak protein-ligand interactions (see Chapter 5) or flexible regions of proteins for which no NOEs can be detected.

1.4 Project Goals

The results obtained during this PhD project are presented in the following four chapters.

As described above, the folding of repeat proteins is an interesting topic to study. We evaluated local folding determinants for consensus designed ankyrin repeat proteins using a range of NMR, biophysical and computational experiments. In order to achieve this we had to first assign the backbone resonances of each repeat and the N- and C-terminal capping repeats – a problem which becomes progressively more difficult as additional internal repeat are added. Using traditional proton-deuterium exchange methods in the presence and absence of chemical denaturation, we evaluated the stability of this ankyrin scaffold in a residue-resolved manner. Rates of rapidly exchanging protons were measured using MEXICO-based (*measurement of fast proton exchange rates in isotopically labelled compounds*) NMR experiments. Protein folding and stability data were analysed in the context of *in silico* predictions based on an Ising-type model. This work is described in the publication included as Chapter 2 of this thesis.

In the ongoing design of a stable consensus Armadillo repeat protein scaffold for use in the directed evolution of peptide binders and building on the work of Parmeggiani *et al.*, a molecular dynamics approach supported by NMR experiments has been employed to refine and stabilise the N- and C-terminal capping repeats as well as the internal M-type repeat. The empirical evaluation of high-ranking stabilising mutations suggested by *in silico* approaches was an important step in validating protein design improvements. In Chapter 3 we detail our efforts in determining these stabilising mutations by MD and the assessment of the effects of these mutations on cap and overall protein stability in the context of a four-repeat consensus designed Armadillo repeat protein, using a range of biophysical and NMR-based methods.

During the course of our work on armadillo repeat proteins a novel approach utilising the previously unknown property of consensus designed armadillo fragments to self-assemble was discovered. This knowledge has been implemented to aid in the resonance assignment of complicated repeat proteins. The proof of principle for this approach, presented in Chapter 4, was investigated with the aim to structurally determine the self-assembly complex of the reconstituted and, where possible, free armadillo repeat protein fragments.

Constituting the largest part of the work undertaken during this thesis, and following up on the scaffold stabilisation work covered in Chapter 3, the characterisation of a selected peptide binder, VG_328 and the nature of its interaction with the neurotensin peptide has been undertaken. Due to the low affinity of the peptide-protein interaction, NMR

experiments were used extensively in the characterisation of this interaction. The larger number of armadillo repeats in the selected 5-repeat VG_328 represented a severe technical challenge for detailed NMR analysis, and strategies to overcome this challenge have been devised. Using a range of NMR-based techniques including fragment-based isotopic labelling, paramagnetic relaxation enhancement, protein truncation, and chemical shift perturbation mapping, extensive resonance assignments were achieved. Experimentally obtained data were combined with molecular dynamics simulations to subsequently develop a model of the binding mode. This work was complemented by further mutational studies involving N-cap mutations and their effects on peptide binding and protein stability. The results of this project are presented in Chapter 5.

While the results presented in Chapters 2 and 3 have been published in their entirety, results presented in Chapters 4 and 5 are part of ongoing research efforts.

1.5 References

1. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats. structures, functions, and evolution. *Journal of structural biology* **134**, 117–131.
2. Grove, T. Z., Cortajarena, A. L. & Regan, L. (2008). Ligand binding by repeat proteins. natural and designed. *Current opinion in structural biology* **18**, 507–515.
3. Kajava, A. V. (2001). Review: proteins with repeated sequence--structural prediction and modeling. *J. Struct. Biol.* **134**, 132–144.
4. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.* **293**, 151–160.
5. Main, E. R., Lowe, A. R., Mochrie, S. G., Jackson, S. E. & Regan, L. (2005). A recurring theme in protein engineering. the design, stability and folding of repeat proteins. *Current opinion in structural biology* **15**, 464–471.
6. Kobe, B. & Kajava, A. V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.* **25**, 509–515.
7. Groves, M. R. & Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9**, 383–389.
8. Stirnimann, C. U., Petsalaki, E., Russell, R. B. & Muller, C. W. (2010). WD40 proteins propel cellular networks. *Trends in biochemical sciences* **35**, 565–574.
9. Cortajarena, A. L., Mochrie, S. G. J. & Regan, L. (2008). Mapping the energy landscape of repeat proteins using NMR-detected hydrogen exchange. *J. Mol. Biol.* **379**, 617–626.
10. Main, E. R. G., Jackson, S. E. & Regan, L. (2003). The folding and design of repeat proteins: reaching a consensus. *Curr. Opin. Struct. Biol.* **13**, 482–489.
11. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Muller, C. W. & Bork, P. (2001). Comparison of ARM and HEAT protein repeats. *Journal of molecular biology* **309**, 1–18.
12. Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K. & Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *Journal of molecular biology* **376**, 241–257.

13. Heringa, J. (1998). Detection of internal repeats: how common are they? *Curr. Opin. Struct. Biol.* **8**, 338–345.
14. Mosavi, L. K., Cammett, T. J., Desrosiers, D. C. & Peng, Z.-Y. (2004). The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci.* **13**, 1435–1448.
15. Björklund, A. K., Ekman, D. & Elofsson, A. (2006). Expansion of protein domain repeats. *PLoS Comput. Biol.* **2**, e114.
16. Bennett, V. & Stenbuck, P. J. (1979). Identification and partial purification of ankyrin, the high affinity membrane attachment site for human erythrocyte spectrin. *J. Biol. Chem.* **254**, 2533–2541.
17. Bennett, V. & Stenbuck, P. J. (1979). The membrane attachment protein for spectrin is associated with band 3 in human erythrocyte membranes. *Nature* **280**, 468–473.
18. Lux, S. E., Tse, W. T., Menninger, J. C., John, K. M., Harris, P., Shalev, O. *et al. (*1990). Hereditary spherocytosis associated with deletion of human erythrocyte ankyrin gene on chromosome 8. *Nature* **345**, 736–739.
19. Sedgwick, S. G. & Smerdon, S. J. (1999). The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem. Sci.* **24**, 311–316.
20. Breeden, L. & Nasmyth, K. (1987). Similarity between cell-cycle genes of budding yeast and fission yeast and the Notch gene of Drosophila. *Nature* **329**, 651–654.
21. Mosavi, L. K., Minor, D. L. & Peng, Z.-Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 16029–16034.
22. Forrer, P., Stumpp, M. T., Binz, H. K. & Plückthun, A. (2003). A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett.* **539**, 2–6.
23. Nüsslein-Volhard, C. & Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795–801.
24. Wieschaus, E., Nusslein-Volhard, C. & Kluding, H. (1984). Krüppel, a gene whose activity is required early in the zygotic genome for normal embryonic segmentation. *Dev. Biol.* **104**, 172–186.
25. Eaton, S. & Cohen, S. (1996). Wnt signal transduction: more than one wat to skin a (beta-)cat? *Trends Cell Biol.* **6**, 287–290.
26. Riggleman, B., Wieschaus, E. & Schedl, P. (1989). Molecular analysis of the armadillo locus. uniformly distributed transcripts and a protein with novel internal repeats are associated with a Drosophila segment polarity gene. *Genes & development* **3**, 96–113.
27. Peifer, M., Berg, S. & Reynolds, A. B. (1994). A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell* **76**, 789–791.
28. Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). Three-dimensional structure of the armadillo repeat region of beta-catenin. *Cell* **90**, 871–882.
29. Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. (1998). Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. *Cell* **94**, 193–204.
30. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368.
31. Main, E. R. G., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. (2003). Design of

stable alpha-helical arrays from an idealized TPR motif. *Structure* **11**, 497–508.
32. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332**, 489–503.
33. Richardson, J. S. & Richardson, D. C. (1989). The de novo design of protein structures. *Trends Biochem. Sci.* **14**, 304–309.
34. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L. *et al. (*2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329**, 309–313.
35. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A. *et al. (*2008). De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391.
36. Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J. *et al. (*2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195.
37. Hanes, J., Schaffitzel, C., Knappik, A. & Plückthun, A. (2000). Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. *Nat. Biotechnol.* **18**, 1287–1292.
38. McCafferty, J., Griffiths, A. D., Winter, G. & Chiswell, D. J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552–554.
39. Bratkovic, T. (2010). Progress in phage display: evolution of the technique and its application. *Cell. Mol. Life Sci.* **67**, 749–767.
40. Pepper, L. R., Cho, Y. K., Boder, E. T. & Shusta, E. V. (2008). A decade of yeast surface display technology: where are we now? *Comb. Chem. High Throughput Screen.* **11**, 127–134.
41. Virnekäs, B., Ge, L., Plückthun, A., Schneider, K. C., Wellnhofer, G. & Moroney, S. E. (1994). Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res.* **22**, 5600–5607.
42. Zaccolo, M., Williams, D. M., Brown, D. M. & Gherardi, E. (1996). An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J. Mol. Biol.* **255**, 589–603.
43. Zhao, H., Giver, L., Shao, Z., Affholter, J. A. & Arnold, F. H. (1998). Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat. Biotechnol.* **16**, 258–261.
44. Stemmer, W. P. (1994). Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389–391.
45. Jermutus, L., Honegger, A., Schwesinger, F., Hanes, J. & Plückthun, A. (2001). Tailoring in vitro evolution for protein affinity or stability. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 75–80.
46. Plückthun, A. (2012). Ribosome display: a perspective. *Methods Mol. Biol.* **805**, 3–28.
47. Interlandi, G., Wetzel, S. K., Settanni, G., Plückthun, A. & Caflisch, A. (2008). Characterization and further stabilization of designed ankyrin repeat proteins by combining molecular dynamics simulations and experiments. *Journal of molecular biology* **375**, 837–854.
48. Merz, T., Wetzel, S. K., Firbank, S., Plückthun, A., Grütter, M. G. & Mittl, P. R. E. (2008). Stabilizing ionic interactions in a full-consensus ankyrin repeat protein. *J.*

Mol. Biol. **376**, 232–240.
49. Binz, H. K., Amstutz, P., Kohl, A., Stumpp, M. T., Briand, C., Forrer, P. *et al. (*2004). High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol.* **22**, 575–582.
50. Stahl, A., Stumpp, M. T., Schlegel, A., Ekawardhani, S., Lehrling, C., Martin, G. *et al. (*2013). Highly potent VEGF-A-antagonistic DARPins as anti-angiogenic agents for topical and intravitreal applications. *Angiogenesis* **16**, 101–111.
51. Kohl, A., Binz, H. K., Forrer, P., Stumpp, M. T., Plückthun, A. & Grütter, M. G. (2003). Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1700–1705.
52. Binz, H. K. & Plückthun, A. (2005). Engineered proteins as specific binding reagents. *Current opinion in biotechnology* **16**, 459–469.
53. Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* **23**, 1105–1116.
54. Gebauer, M. & Skerra, A. (2009). Engineered protein scaffolds as next-generation antibody therapeutics. *Curr Opin Chem Biol* **13**, 245–255.
55. Skerra, A. (2008). Alternative binding proteins: anticalins - harnessing the structural plasticity of the lipocalin ligand pocket to engineer novel binding activities. *FEBS J.* **275**, 2677–2683.
56. Stanfield, R. L. & Wilson, I. A. (1995). Protein-peptide interactions. *Current opinion in structural biology* **5**, 103–113.
57. MacCallum, R. M., Martin, A. C. & Thornton, J. M. (1996). Antibody-antigen interactions. contact analysis and binding site topography. *Journal of molecular biology* **262**, 732–745.
58. Almagro, J. C. (2004). Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size. implications for the rational design of antibody repertoires. *Journal of molecular recognition : JMR* **17**, 132–143.
59. Sundberg, E. J. (2009). Structural basis of antibody-antigen interactions. *Methods in molecular biology* **524**, 23–36.
60. Fabio Parmeggiani (2008). Design of Armadillo Repeat Protein Scaffolds. PhD, Switzerland.
61. Kuriyan, J. & Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Annual review of biophysics and biomolecular structure* **26**, 259–288.
62. Pawson, T. & Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445–452.
63. Pawson, T. & Scott, J. D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278**, 2075–2080.
64. Neefjes, J., Jongsma, M. L., Paul, P. & Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature reviews. Immunology* **11**, 823–836.
65. Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. (2006). How TCRs bind MHCs, peptides, and coreceptors. *Annual review of immunology* **24**, 419–466.
66. Vyas, J. M., van der Veen, A. G. & Ploegh, H. L. (2008). The known unknowns of antigen processing and presentation. *Nature reviews. Immunology* **8**, 607–618.
67. Esteban, O. & Zhao, H. (2004). Directed evolution of soluble single-chain human

class II MHC molecules. *J. Mol. Biol.* **340**, 81–95.
68. Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. (1999). The WD repeat. a common architecture for diverse functions. *Trends in biochemical sciences* **24**, 181–185.
69. Blatch, G. L. & Lassle, M. (1999). The tetratricopeptide repeat. a structural motif mediating protein-protein interactions. *BioEssays : news and reviews in molecular, cellular and developmental biology* **21**, 932–939.
70. Andrade, M. A. & Bork, P. (1995). HEAT repeats in the Huntington's disease protein. *Nat. Genet.* **11**, 115–116.
71. Kippert, F. & Gerloff, D. L. (2009). Highly sensitive detection of individual HEAT and ARM repeats with HHpred and COACH. *PloS one* **4**, e7148.
72. Coates, J. C. (2003). Armadillo repeat proteins. beyond the animal kingdom. *Trends in cell biology* **13**, 463–471.
73. Chen, C. K., Chan, N. L. & Wang, A. H. (2011). The many blades of the beta-propeller proteins. conserved but versatile. *Trends in biochemical sciences* **36**, 553–561.
74. Nikkhah, M., Jawad-Alami, Z., Demydchuk, M., Ribbons, D. & Paoli, M. (2006). Engineering of beta-propeller protein scaffolds by multiple gene duplication and fusion of an idealized WD repeat. *Biomolecular engineering* **23**, 185–194.
75. D'Andrea, L. D. & Regan, L. (2003). TPR proteins. the versatile helix. *Trends in biochemical sciences* **28**, 655–662.
76. Cortajarena, A. L., Yi, F. & Regan, L. (2008). Designed TPR modules as novel anticancer agents. *ACS chemical biology* **3**, 161–166.
77. Cortajarena, A. L., Liu, T. Y., Hochstrasser, M. & Regan, L. (2010). Designed proteins to modulate cellular networks. *ACS chemical biology* **5**, 545–552.
78. Cortajarena, A. L., Wang, J. & Regan, L. (2010). Crystal structure of a designed tetratricopeptide repeat module in complex with its peptide ligand. *The FEBS journal* **277**, 1058–1066.
79. Urvoas, A., Guellouz, A., Valerio-Lepiniec, M., Graille, M., Durand, D., Desravines, D. C. *et al. (*2010). Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (alphaRep) based on thermostable HEAT-like repeats. *Journal of molecular biology* **404**, 307–327.
80. Hatzfeld, M. (1999). The armadillo family of structural proteins. *International review of cytology* **186**, 179–224.
81. Tewari, R., Bailes, E., Bunting, K. A. & Coates, J. C. (2010). Armadillo-repeat protein functions. questions for little creatures. *Trends in cell biology* **20**, 470–481.
82. Catimel, B., Teh, T., Fontes, M. R., Jennings, I. G., Jans, D. A., Howlett, G. J. *et al. (*2001). Biophysical characterization of interactions involving importin-alpha during nuclear import. *The Journal of biological chemistry* **276**, 34189–34198.
83. Reichen, C., Hansen, S. & Plüchthun, A. (2013). Modular Peptide Binding: From a comparison of natural binders to designed Armadillo Repeat Proteins, accepted.
84. Conti, E. & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure* **8**, 329–338.
85. Fontes, M. R., Teh, T., Jans, D., Brinkworth, R. I. & Kobe, B. (2003). Structural basis for the specificity of bipartite nuclear localization sequence binding by importin-alpha. *The Journal of biological chemistry* **278**, 27981–27987.
86. Kobe, B. (1999). Autoinhibition by an internal nuclear localization signal revealed by the crystal structure of mammalian importin alpha. *Nature structural biology* **6**,

388–397.
87. Pumroy, R. A., Nardozzi, J. D., Hart, D. J., Root, M. J. & Cingolani, G. (2012). Nucleoporin Nup50 stabilizes closed conformation of armadillo repeat 10 in importin alpha5. *The Journal of biological chemistry* **287**, 2022–2031.
88. Madhurantakam, C., Varadamsetty, G., Grütter, M. G., Plückthun, A. & Mittl, P. R. (2012). Structure-based optimization of designed Armadillo-repeat proteins. *Protein science : a publication of the Protein Society* **21**, 1015–1028.
89. Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-Sato, E., Tomita, M. *et al. (*2009). Six classes of nuclear localization signals specific to different binding grooves of importin alpha. *The Journal of biological chemistry* **284**, 478–485.
90. Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O. *et al. (*2008). Designed armadillo repeat proteins as general peptide-binding scaffolds. consensus design and computational optimization of the hydrophobic core. *Journal of molecular biology* **376**, 1282–1304.
91. Alfarano, P., Varadamsetty, G., Ewald, C., Parmeggiani, F., Pellarin, R., Zerbe, O. *et al. (*2012). Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy. *Protein science : a publication of the Protein Society* **21**, 1298–1314.
92. Kramer, M. A., Wetzel, S. K., Plückthun, A., Mittl, P. R. & Grütter, M. G. (2010). Structural determinants for improved stability of designed ankyrin repeat proteins with a redesigned C-capping module. *Journal of molecular biology* **404**, 381–391.
93. Kajander, T., Cortajarena, A. L., Main, E. R., Mochrie, S. G. & Regan, L. (2005). A new folding paradigm for repeat proteins. *Journal of the American Chemical Society* **127**, 10188–10190.
94. Tripp, K. W. & Barrick, D. (2007). Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. *Journal of molecular biology* **365**, 1187–1200.
95. Varadamsetty, G., Tremmel, D., Hansen, S., Parmeggiani, F. & Plückthun, A. (2012). Designed Armadillo repeat proteins. library generation, characterization and selection of peptide binders with high specificity. *Journal of molecular biology* **424**, 68–87.
96. Wetzel, S. K. (2008). Folding and unfolding mechanism of designed ankyrin repeat proteins. Diss. Univ. Zürich, 2008. - Ref.: Andreas Plückthun, Zürich.
97. Nieto, J. L., Rico, M., Santoro, J., Herranz, J. & Bermejo, F. J. (1986). Assignment and conformation of neurotensin in aqueous solution by 1H NMR. *International journal of peptide and protein research* **28**, 315–323.
98. Lowe, A. R. & Itzhaki, L. S. (2007). Biophysical characterisation of the small ankyrin repeat protein myotrophin. *J. Mol. Biol.* **365**, 1245–1255.
99. Lowe, A. R. & Itzhaki, L. S. (2007). Rational redesign of the folding pathway of a modular protein. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 2679–2684.
100. Tang, K. S., Guralnick, B. J., Wang, W. K., Fersht, A. R. & Itzhaki, L. S. (1999). Stability and folding of the tumour suppressor protein p16. *J. Mol. Biol.* **285**, 1869–1886.
101. Tang, K. S., Fersht, A. R. & Itzhaki, L. S. (2003). Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure* **11**, 67–73.
102. Zweifel, M. E. & Barrick, D. (2001). Studies of the ankyrin repeats of the

Drosophila melanogaster Notch receptor. 2. Solution stability and cooperativity of unfolding. *Biochemistry* **40**, 14357–14367.
103. Phillips, J. J., Javadi, Y., Millership, C. & Main, E. R. G. (2012). Modulation of the multistate folding of designed TPR proteins through intrinsic and extrinsic factors. *Protein Sci.* **21**, 327–338.
104. Cortajarena, A. L., Mochrie, S. G. J. & Regan, L. (2011). Modulating repeat protein stability: the effect of individual helix stability on the collective behavior of the ensemble. *Protein Sci.* **20**, 1042–1047.
105. Bieri, M., Kwan, A. H., Mobli, M., King, G. F., Mackay, J. P. & Gooley, P. R. (2011). Macromolecular NMR spectroscopy for the non-spectroscopist: beyond macromolecular solution structure determination. *FEBS J.* **278**, 704–715.
106. Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F. & Mackay, J. P. (2011). Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS J.* **278**, 687–703.
107. Bodenhausen, G. & Ruben, D. J. (1980). Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chemical Physics Letters* **69**, 185–189.
108. Grzesiek, S., Döbeli, H., Gentz, R., Garotta, G., Labhardt, A. M. & Bax, A. (1992). 1H, 13C, and 15N NMR backbone assignments and secondary structure of human interferon-gamma. *Biochemistry* **31**, 8180–8190.
109. Clubb, R. T. & Wagner, G. (1992). A triple-resonance pulse scheme for selectively correlating amide 1HN and 15N nuclei with the 1H alpha proton of the preceding residue. *J. Biomol. NMR* **2**, 389–394.
110. Weisemann, R., Rüterjans, H. & Bermel, W. (1993). 3D triple-resonance NMR techniques for the sequential assignment of NH and 15N resonances in 15N- and 13C-labelled proteins. *J. Biomol. NMR* **3**, 113–120.
111. Fabian Bumbak (2011). Solution Structure of a C-terminal Fragment of a Designed Split Armadillo Repeat Protein, Switzerland.
112. Iwai, H., Züger, S., Jin, J. & Tam, P.-H. (2006). Highly efficient protein trans-splicing by a naturally split DnaE intein from Nostoc punctiforme. *FEBS Lett.* **580**, 1853–1858.
113. Züger, S. & Iwai, H. (2005). Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. *Nat. Biotechnol.* **23**, 736–740.
114. Skrisovska, L. & Allain, F. H.-T. (2008). Improved segmental isotope labeling methods for the NMR study of multidomain or large proteins: application to the RRMs of Npl3p and hnRNP L. *J. Mol. Biol.* **375**, 151–164.
115. Skrisovska, L., Schubert, M. & Allain, F. H.-T. (2010). Recent advances in segmental isotope labeling of proteins: NMR applications to large proteins and glycoproteins. *J. Biomol. NMR* **46**, 51–65.
116. Slynko, V., Schubert, M., Numao, S., Kowarik, M., Aebi, M. & Allain, F. H.-T. (2009). NMR structure determination of a segmentally labeled glycoprotein using in vitro glycosylation. *J. Am. Chem. Soc.* **131**, 1274–1281.
117. Overhauser, A. W. P. R. 9. (. 4. (1953). Polarization of Nuclei in Metals. *Physical Review Letters* **92**, 411–415.
118. Carver, T. & Slichter, C. (1953). Polarization of Nuclear Spins in Metals. *Phys. Rev.* **92**, 212–213.
119. Anderson, W. A. & Freeman, R. (1962). Influence of a Second Radiofrequency

Field on High-Resolution Nuclear Magnetic Resonance Spectra. *Journal of Chemical Physics* **37**, 411–415.

120. Anet, F. A. L. & Bourn, A. J. R. (1965). Nuclear Magnetic Resonance Spectral Assignments from Nuclear Overhauser Effects. *J. Am. Chem. Soc.* **87**, 5250–5251.

121. Jahnke, W. (2002). Spin labels as a tool to identify and characterize protein-ligand interactions by NMR spectroscopy. *Chembiochem* **3**, 167–173.

122. Battiste, J. L. & Wagner, G. (2000). Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. *Biochemistry* **39**, 5355–5365.

123. Bertini, I., Luchinat, C. & Parigi, G. (2002). Paramagnetic constraints: An aid for quick solution structure determination of paramagnetic metalloproteins. *Concepts Magn. Reson.* **14**, 259–286.

124. Wittekind, M. & Mueller, L. (1993). HNCACB, a High-Sensitivy 3D NMR Experiment to Correlate Amide-Proton and Nitrogen Resonances with the Alpha- and Beta-Carbon Resonances in Proteins. *J. Magn. Reson., Series B.* **B101**: 201-205.

125. McCallum, S. A., Hitchens, T. K. & Rule, G. S. (1998). Unambiguous correlations of backbone amide and aliphatic gamma resonances in deuterated proteins. *J. Magn. Reson.* **134**, 350–354

126 Ikura, M., Kay, L. E. and Bax, A. (1990). A novel approach for sequential assignment of proton, carbon-13, and nitrogen-15 spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry* **29**: 4659-4667.

127 Yamazaki, T., Lee, W., Arrowsmith, C.H., Muhandiram, D.R. and Kay, L.E. (1994). A Suite of Triple Resonance NMR Experiments for the Backbone Assignment of ^{15}N, ^{13}C, ^{2}H Labeled Proteins with High Sensitivity. *J. Am. Chem. Soc.* **116**: 11655-11666.

128 Schlinkmann, K. M., Hillenbrand, M., Rittner, A., Künz, M., Strohner, R., Plückthun, A. (2012). Maximizing Detergent Stability and Functional Expression of a GPCR by Exhaustive Recombination and Evolution. J. Mol. Biol. **422**: 414-428

2. Residue-resolved stability of full-consensus ankyrin repeat proteins probed by NMR

Svava K. Wetzel[1,2†], Christina Ewald[2†], Giovanni Settanni[3], Simon Jurt[2], Andreas Plückthun[1]*, Oliver Zerbe[2]*

Published in:
Journal of Molecular Biology, September 2010, Volume 10, issue 402(1), pages 241-58

[1]Institute of Biochemistry, University of Zürich, CH-8057 Zürich, Switzerland
[2]Institute of Organic Chemistry, University of Zürich, CH-8057 Zürich, Switzerland
[3]MRC – Centre for Protein Engineering, Cambridge CB2 0QH, United Kingdom
* Corresponding authors: plueckthun@bioc.uzh.ch; oliver.zerbe@oci.uzh.ch
Corresponding Author Addresses:

Oliver Zerbe
University of Zürich
Institute of Organic Chemistry
Winterthurerstrasse 190
8057-Zurich, Switzerland
Tel: + 41 1 635 42 63
Fax: + 41 1 635 68 84
Email: oliver.zerbe@oci.uzh.ch

Andreas Plückthun
University of Zürich
Department of Biochemistry
Winterthurerstrasse 190
8057-Zurich, Switzerland
Tel: + 41 1 635 55 70
Fax: + 41 1 635 57 12
Email: plueckthun@bioc.uzh.ch

† these authors contributed equally to the work

Abstract

We investigated the stability determinants and the unfolding characteristics of full-consensus Designed Ankyrin Repeat Proteins (DARPins) by NMR. Despite the repeating sequence motifs, the resonances could be fully assigned using ^{1}H, ^{15}N, ^{13}C triple labeled proteins. To remove further ambiguities, paramagnetic spin labels were attached to either end of these elongated proteins which attenuate the resonances of the

spatially closest residues. Deuterium exchange experiments of DARPins with 2 and 3 internal repeats between N- and C-terminal capping repeats (NI_2C, NI_3C) and NI_3C_Mut5, where the C-cap had been reengineered, indicate that the stability of the full-consensus ankyrin repeat proteins is strongly dependent on the coupling between repeats, as the stabilized cap decreases the exchange rate throughout the whole protein. Some amide protons require more than a year to exchange at 37 °C, highlighting the extraordinary stability of the proteins. Denaturant induced unfolding, followed by deuterium exchange, chemical shift change and heteronuclear nuclear Overhauser effects, is consistent with an Ising-type description of equilibrium folding for NI_3C_Mut5, while for native state deuterium exchange, we postulate local fluctuations to dominate exchange as unfolding events are too slow in these very stable proteins. The location of extraordinarily slowly exchanging protons indicate a very stable core structure in the DARPins which combines hydrophobic shielding with favorable electrostatic interactions. These investigations help the understanding of repeat protein architecture and the further design of DARPins for biomedical applications where high stability is required.

2.2 Introduction

Repeat proteins are built of repeating structural units of typically 25-45 amino acids that stack together to build a folded domain.[1, 2] Amongst the most common types of repeat motifs are the ankyrin repeat, armadillo repeat, leucine-rich repeat and tetratricopeptide repeat.[3]

The repeat protein architecture relies on stabilizing and structure-determining interactions formed within a repeat and between the neighboring repeats, and contains no interactions between residues very far apart in the protein sequence. This modular nature of repeat proteins makes them fundamentally different from globular proteins, and thus interesting for testing experimental and theoretical views that have emerged from the study of globular proteins.

The ankyrin repeat (AR) is a 33-residue motif consisting of a β-turn, followed by two antiparallel α-helices and a loop reaching towards the turn of the next repeat.[4] A library of Designed Ankyrin Repeat Proteins (DARPins)[5] has been created as a source of very robust specific binding proteins for many applications in biomedicine and biochemical research. It has thus been interesting both from a fundamental and an applied perspective to understand the stability determinants of these proteins. The library contains internal repeats with randomized residues flanked by an N- and a C-terminal capping repeat.[6] The capping repeats are essential to allow the folding of these proteins within the cell,[7] and these proteins can be expressed in soluble form to very high levels.

To systematically investigate the stability determinants, we have previously designed a full-consensus AR as an idealized example for studying AR protein folding and constructed DARPins of variable length, built up by combining the N- and C-capping module (N, C) with internal identical repeats of varying number (I_1 to I_6).[8] By a systematic study of these proteins we found an increase of stability with length, up to the point that the proteins became resistant to boiling and saturated GdmCl solutions.[8]

We have previously investigated three of our designed proteins, NI_1C, NI_2C and NI_3C, not only by equilibrium, but also kinetic unfolding and refolding analysis, and found that all three proteins display a complex folding mechanism. While they all revealed at least a three-state mechanism in kinetic experiments, NI_3C demonstrated a stable intermediate state that is also detected in equilibrium CD measurements.[8] We suggested a possible structure of this intermediate, where the N-terminal capping repeat and all three internal full-consensus repeats are still folded, while the C-terminal capping repeat is unfolded.[7]

Combining Molecular Dynamics simulations and experiments,[7] we deduced that the C-terminal capping repeat, taken from a natural AR protein[6] was the weakest link, and designed improved C-caps with much higher stability.[7] From GdmCl equilibrium unfolding experiments, these new capping repeats appear to stabilize the whole protein, as the main transition is shifted to even higher GdmCl concentrations.

We wished to now obtain more detailed information of the stability determinants of these extraordinarily denaturation-resistant repeat proteins and of the structure of their intermediates and therefore used hydrogen/deuterium exchange (HX) observed through 2D ^{1}H-^{15}N correlation spectroscopy.[9] The hydrogen exchange reactions of amides in the native state of proteins are the result of structural fluctuations that can be divided into

three categories: (i) cooperative global unfolding (all residues), (ii) cooperative local unfolding (several residues) and (iii) non-cooperative local fluctuations of individual residues.

In order to relate the equilibrium stability of repeat proteins to the number of repeats a 1D Ising-like model was used to describe the transitions.[8,10-12] This model could also describe the unfolding kinetics as a function of repeat number.[8] According to this model, each repeat is considered to be an independent folding unit that interacts with the neighboring repeat(s). In the case of DARPins five parameters could describe the system: the coupling free energy *J* between repeats, the free energy of an isolated repeat ΔG_0, its denaturant dependence *m*, the free energy of an isolated capping repeat $\Delta G'_0$ and its denaturant dependence *m'*.[8]

In the present study, we intended to analyze the local stability of three proteins, NI_2C, NI_3C and NI_3C with the more stable C-cap, termed NI_3C_Mut5 (for sequences see Figure S2.1). We present an approach for deriving nearly complete backbone assignments, a non-trivial task due to the highly repetitive nature of the sequences. Amide proton exchange is measured using classical $^1H/^2H$ exchange experiments both in the absence and presence of denaturant. In order to capture the rates of the more rapidly exchanging residues, MEXICO (*measurement of fast proton exchange rates in isotopically labeled compounds*) experiments[13] were additionally performed. Taken together, we could obtain an extensive set of protection factors along the sequence spanning 9 orders of magnitude. Furthermore, we monitored GdmCl-induced equilibrium unfolding by using ^1H-^{15}N heteronuclear NMR techniques. The data are discussed within the context of folding models, in particular with respect to Ising-type folding models.

2.3 Results

2.3.1 Backbone assignment of NI2C, NI3C and NI3C_Mut5

As none of the DARPins has been studied by NMR so far, the first step was to assign the ^{1}H,^{13}C,^{15}N resonances of NI_2C, NI_3C and NI_3C_Mut5 using 3D triple-resonance NMR experiments and paramagnetic relaxation enhancement (PRE) spin labels. Signal dispersion was very good despite the α-helical fold of the proteins and the presence of identical sequence fragments (Figure 2.1 and S2.2Figure S2.3 and Figure S2.4), and line widths indicate that all proteins are monomeric species.

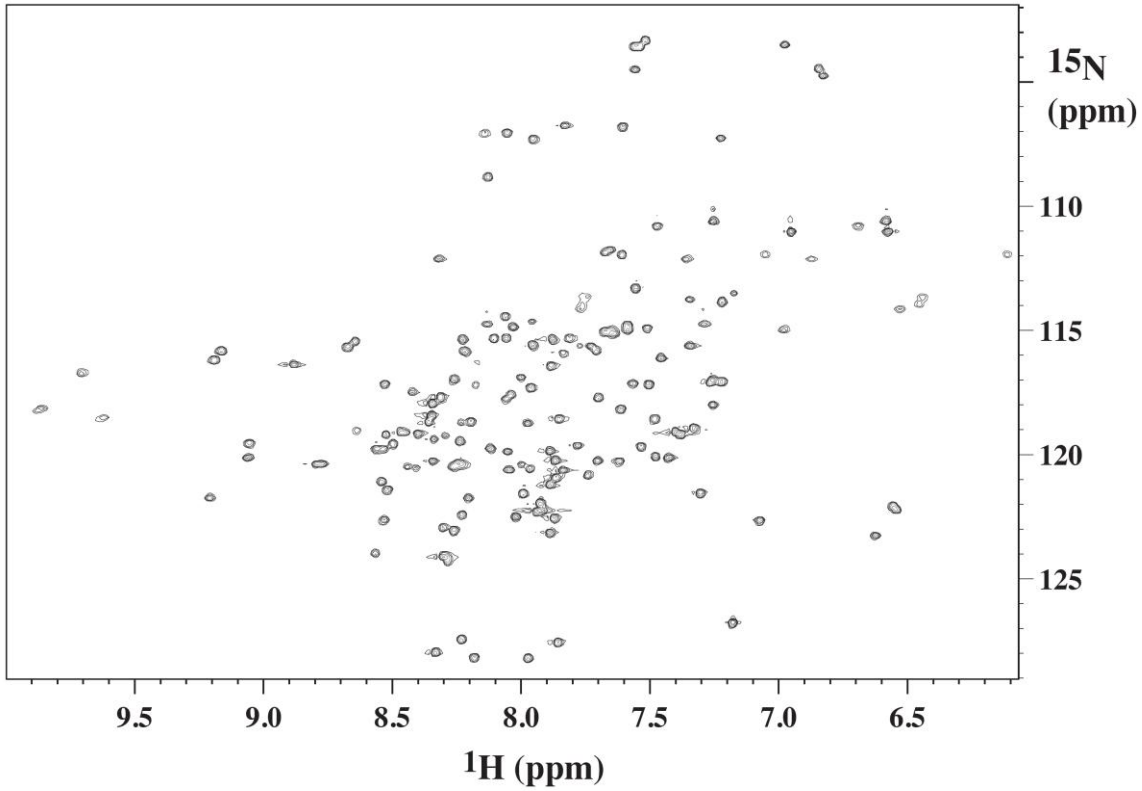


Figure 2.1 700 MHz [^{15}N, ^{1}H]- HSQC spectrum of 1.5 mM ^{15}N,^{13}C,^{2}H-labeled NI_3C_Mut5 in 50 mM phosphate, 150 mM NaCl, pH 7.4 at 310 K. For spectra of NI_2C, NI_3C and assignments, see Supplementary Material.

The [^{15}N,^{1}H]-HSQC spectra of NI_2C and NI_3C with the original C-cap display peaks of a second, much more flexible conformation at the C-terminus, which is absent in NI_3C_Mut5 (*vide infra*). Peak volume ratios of the predominant conformation to this second conformation are 2-3.5 in case of NI_2C and 1-2.7 in case of NI_3C. Due to differences in T2 relaxation times of residues in the folded versus the unfolded conformation these ratios cannot be reliably translated into population ratios, as the unfolded population would be overemphasized. The occurrence of a second minor species with the last 5 C-terminal residues unfolded, probably in equilibrium with the native species, is consistent with the previously observed more facile denaturation of the original C-cap.[7, 8]

The full-consensus DARPins investigated here contain two or three identical internal repeats. Due to the near-identical structure of the individual repeats, residues at the corresponding position in different repeats possess very similar chemical shifts. However, the fact that in almost all cases peaks were individually resolved in the [^{15}N,^{1}H]-HSQC spectra prompted us to exclusively use ^{2}H,^{13}C,^{15}N-labeled proteins with constant-time 3D out-and-back experiments that terminate with recording amide proton signals. Since carbon resonances of corresponding positions in the individual repeats are extremely similar, backbone assignment must take advantage of the improved resolution in the proton-nitrogen correlation map. Critically important for the assignment process was the successful use of the HN(COCA)NH experiment that directly correlates neighboring amide moieties and only requires that resonances are resolved in the ^{15}N dimension (Figure S2.5). 3D HNCO and HN(CA)CO experiments were recorded and additionally utilized for assignments. Finally, the assignments were cross-validated against all 3D spectra, and assignments were only accepted when segments could be successfully traced completely in the HN(COCA)NH, HN(CA)CO and ^{15}N-resolved NOESY spectra.
While most ambiguities could be successfully resolved using this set of experiments, significant difficulties with the backbone assignment for the internal repeats remained. For example, NI_3C and NI_3C_Mut5 both contain two identical fragments spanning the first and second internal repeats, which are flanked by proline residues. While assignments within the repeats were achieved, the resulting lack of information on the sequential connectivity of these fragments made it impossible to distinguish assignments from the individual internal repeats. These difficulties could be successfully resolved by using a nitroxyl spin label (1-oxyl-2,2,5,5-tetramethyl-Δ3-pyrroline-3-methyl)mercaptyl = MTSL), attached by forming a disulfide with unique cysteines, which were introduced into the N- or C-cap. MTSL coupling resulted in the reduction of peak intensities of the respective cap region and parts of the adjacent repeat. Figure 2.2 displays relative intensities of amide moieties for the spin-labeled mutants NI_3C_Mut5-D28C (a) and -D155C (b), respectively (see Figure S2.6 for corresponding data of NI_2C and NI_3C). For example, MTSL coupled to position 155 affects the C-cap and parts of the third repeat. The resulting pattern of distance-dependent signal attenuation in [^{15}N,^{1}H]-HSQC spectra was compared to the theoretical values computed from crystal structures and models (data not shown).

To ensure that the observed signal attenuation was not due to intermolecular effects, the respective isotopically unlabeled Cys-mutant was coupled to MTSL and mixed with the N-ethyl-maleimide (NEM)-protected ^{15}N-labeled Cys-mutant in a 1:2 (^{15}N-NEM:^{14}N-MTSL) ratio. In such an experiment, attenuations can only be caused by intermolecular effects. The comparison of the spectra revealed, however, that no signal reductions due to intermolecular PREs occur (see Figure S2.7).

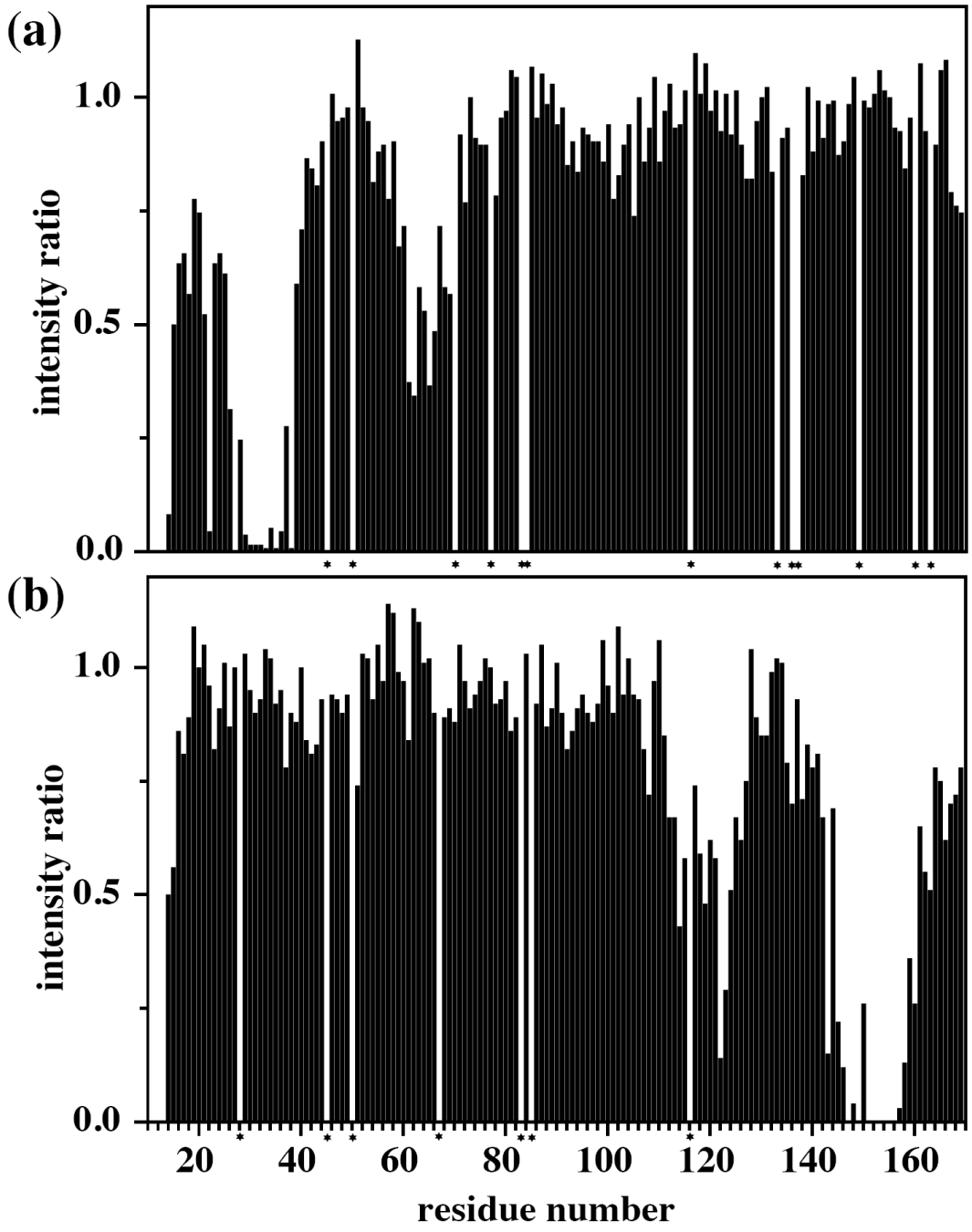


Figure 2.2 Paramagnetic relaxation enhancement (PRE) data of D28C (a) and D155C (b) mutants of NI_3C_Mut5. Bars represent the intensity ratios of cross peaks (MTSL-protein:reference) in the [^{15}N, ^{1}H]- HSQC spectra of the MTSL-labeled Cys-mutant relative to the same mutant in which the Cys residue was reduced with DTT to avoid disulfide formation and not MTSL-labeled. Values for residues that cannot be reliably integrated have been omitted and marked by a star directly below the axis.

NI_2C, NI_3C and NI_3C_Mut5 were assigned by using the procedure described above, starting from NI_2C as the smallest member of this series and exploiting this information for subsequent assignment of NI_3C and NI_3C_Mut5. The overall completeness of the backbone assignment was 99% for NI_2C, NI_3C and NI_3C_Mut5 (see also Figures S2-S4).

The flexibility of the second C-terminal conformation in the wild-type C-cap of NI_2C and NI_3C was verified by using heteronuclear NOE measurements. Values above 0.6 indicate well-structured regions of the protein with little internal flexibility.[14] While the $^{15}N\{^1H\}$-NOE adopted values around 0.8 for most residues and positive values throughout all residues in both proteins, a second set of peaks characterized by negative $^{15}N\{^1H\}$-NOEs was present that was assigned to the C-terminal residues (Figure 2.3). In contrast, no such additional peaks were observed in the [^{15}N,^1H]-HSQC of NI_3C_Mut5.

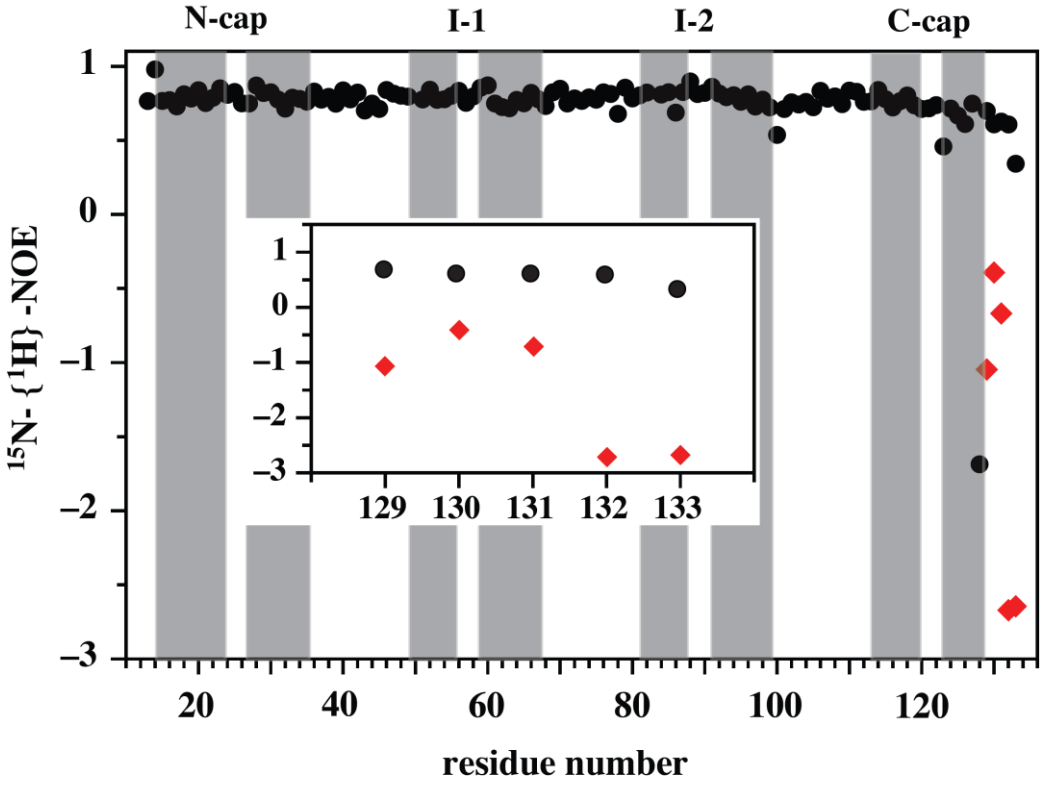


Figure 2.3 $^{15}N\{^1H\}$-NOE data for NI_2C recorded at 600 MHz (black circles). Data corresponding to a C-terminal minor conformation are depicted as red diamonds. The inset shows an expansion of the values at the C terminus.

2.3.2 Residue-resolved stability mapping using amide proton exchange

The full-consensus DARPins are extremely stable and thus will show $^1H/^2H$ exchange rates spanning many orders of magnitude. For this reason we had to use three separate experiments to cover the full range of exchange rates: (i) To measure protons with intermediate exchange rates, lyophilized protein was dissolved in ^2H_2O, and exchange was monitored by recording [^{15}N,^1H]-HSQC spectra at 290 K over 22 h. (ii) To record the extremely slow exchange rates of protons still present after 22 h, another sample of the protein was incubated at 310 K for several months. $^1H/^2H$ exchange was completed for all residues after 3 months at 310 K in case of NI_2C and NI_3C, while for NI_3C_Mut5 even after 12 months 8 amide cross-peaks were still visible (*vide infra*). We therefore destabilized the protein by incubating it at 60°C or by adding GdmCl[15] and re-measured

exchange rates (*vide infra*). (iii) Finally, to also determine the rates of the fast-exchanging residues in the millisecond time range we used the MEXICO experiment.[13] This method "bleaches" all nitrogen-attached protons and monitors re-appearance of signal intensity due to exchange with water protons. Exchange rates of 46 amides of NI_2C, 24 amides of NI_3C and 24 amides of NI_3C_Mut5, which had disappeared before the first HSQC spectrum at 290 K, could be recorded by this method.
The rates for individual amide protons were obtained by fitting the exponential functions to corresponding cross-peak intensities as described in Materials & Methods. Examples of such fits for amides from the internal repeats as well as from the N- or C-caps of NI_3C are depicted in Figure 2.4.

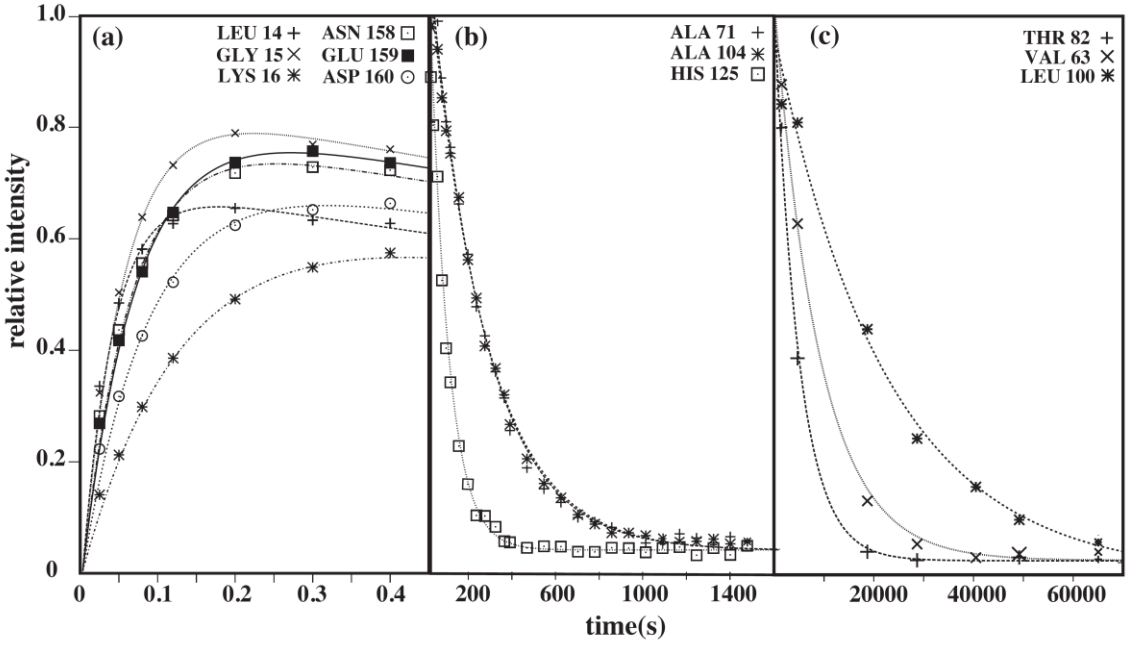


Figure 2.4 Sample signal decay curves of NI_3C in the hydrogen exchange experiments. Curves correspond to data recorded using the MEXICO experiment (a), as well as classical $^1H/^2H$ exchange recorded at 290 K (b) and 310 K (c). Lines correspond to best fits of analytical functions (eq. 1 and 2) as described in the Materials & Methods section.

All rates measured at higher temperature were normalized to 293 K by using the known T-dependence of $^1H/^2H$ exchange.[16] They are plotted as protection factors (PF) as a function of sequence position for NI_2C, NI_3C and NI_3C_Mut5 (Figure 2.5). The PFs can then be converted into the stability difference $\Delta G_{(HX)}$ between the protected and the unprotected state of the amide (see Materials & Methods).

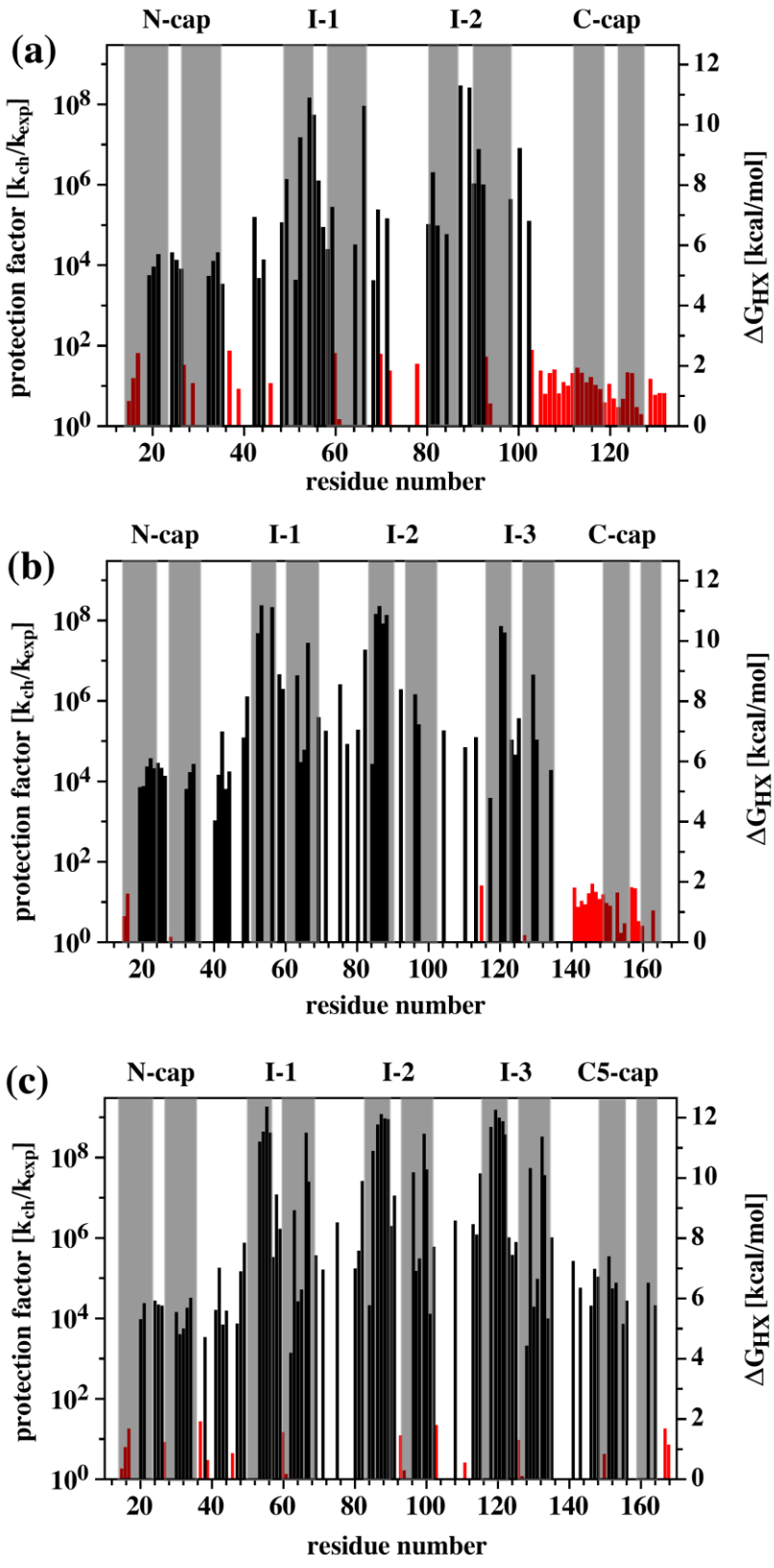


Figure 2.5 Hydrogen exchange data for NI_2C, NI_3C and NI_3C_Mut5. The free energy of exchange ΔG(HX) and the protection factors (PF) translated to 293 K are plotted as a function of residue number for (a) NI_2C, (b) NI_3C, (c) NI_3C_Mut5. The secondary structure is indicated by grey background (helices). The repeat types (N-cap, C-cap, I) are indicated on top of each plot, the PF derived from fast MEXICO rates are shown in red, from the $^1H/^2H$ exchange experiments in black.

A few general features emerge from these investigations of DARPins (Figures 5, 6): (i) Protection factors (PFs) of the capping repeats are generally lower. (ii) The PFs are higher for the N-cap than for the wild-type C-cap, while PFs of the newly designed C-cap of $NI_3C\_Mut5^7$ are comparable to those of the N-cap. The stability of the N-cap as computed from the PF is about 6 kcal/mol in all three DARPins measured, while the stability of the wild-type C-cap is around 1.5 kcal/mol. The redesigned C-cap of NI_3C_Mut5 is about 6.5 kcal/mol and thus even slightly more stable than the N-cap. (iii) Within the internal repeats PFs from residues in helix 1 are generally higher than those in helix 2. (iv) The PFs of NI_3C are uniform within all internal repeats and not higher for the central internal repeat of NI_3C and its mutant NI_3C_Mut5, which will be discussed in the context of expectations from an Ising model below (see Discussion) (v) Exchange rates of some of the residues from the most internal repeat of NI_3C_Mut5 are so low that a few peaks in the spectra can be followed over the time span of more than a year at 37°C. Those residues are generally found at central positions of the helices (e.g. A55, A88, A121 in helix 1 and L66, L99, L132 in helix 2 of repeats 1, 2 and 3 respectively), and these PFs are close to 10^9 (Figure 2.5).

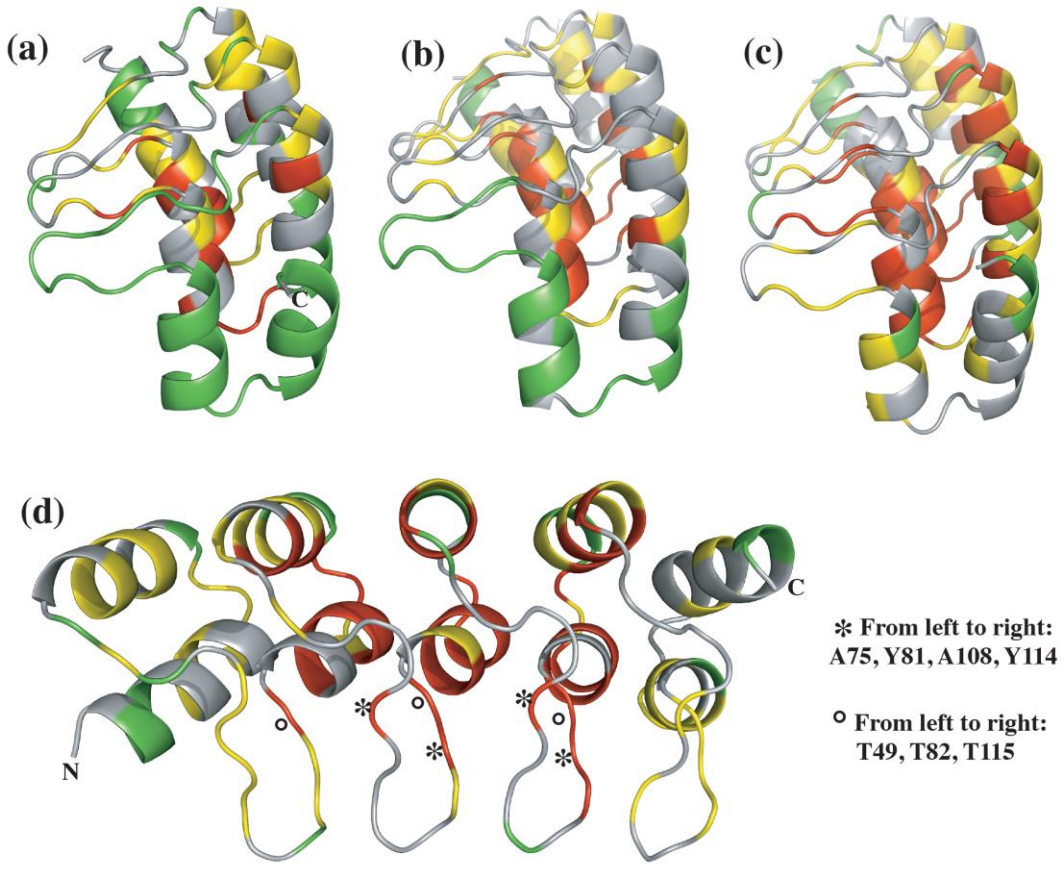


Figure 2.6 Mapping of the protection factors by colour onto the structures of NI_2C (a), NI_3C (b) and NI_3C_Mut5 (c,d), according to the protection factors of the corresponding backbone amide protons at 293 K. Residues with very low protection factors ranging from 10-1 to 102 are colored in green, residues with moderate PF from 103 to 3·105 in yellow and residues with the highest protection in red (PF from 4.5·105 to 109). Residues for which peaks could not be integrated due to overlap are shown in grey.

Besides these general observations, specific structural features become apparent from this amide proton exchange analysis. Several residues located in long loops connecting to the next repeat of NI_3C are surprisingly well protected. For example, residues A75, A108, and Y114, whose backbone carbonyl groups are involved in hydrogen bonds with side chains of H52 and H85, display high PFs (Figure 2.6).

These histidine residues are part of the very conserved TPLH motifs, located at the beginning of the first helix of the internal repeats, which have previously been recognized as forming several hydrogen bonds thus connecting adjacent loops. Conversely, each of these loop is thereby held in place by making contacts to two histidines of TPLH motifs from adjacent repeats.[7, 17]

The rates of the most slowly exchanging amides in NI2C are characterized by PFs of about 3·108, corresponding to a DG(HX) of 11 kcal/mol (Figure 2.5). This value is in approximate agreement with the DG determined previously from denaturant-induced global unfolding (9.2 ± 0.7 kcal/mol)8. Most rates, however, correspond to a lower DG(HX), indicating that exchange processes other than global unfolding, such as local unfolding and fluctuation events, must be taken into account (Figure 2.5(a)). Denaturant-induced global unfolding curves for NI_3C and NI_3C_Mut5 were fitted assuming 3 and 2-state folding (19.7 ± 4.6 kcal/mol8 and 17.9 ± 0.7 kcal/mol, respectively). This fit, however, presents only a crude approximation because it neglects the stability difference of the C- and N-caps and assumes cooperative folding. But even when considering the large error associated with such a fit, those data clearly demonstrate that the DG(HX) obtained in the present work from the slowest exchanging residues (about 11 kcal/mol and 12.5 kcal/mol for NI_3C and NI_3C_Mut5, respectively, (Figure 2.5(b) and Figure 2.5(c))) is much lower than the value from equilibrium denaturant-induced unfolding, and this difference is outside of the range of error. Therefore we conclude that the 1H/2H exchange of the most slowly exchanging protons is not dominated by global unfolding, but contains major contributions from local unfolding and local fluctuation processes. The large difference between the protection factors of the caps and the internal repeats, apparently associated to local unfolding events, emphasizes the non-cooperative nature of the folding process of these proteins, which cannot be fully captured by two- or three-state folding models.
NI_3C_Mut5 with the more stable C-cap has a denaturant-transition midpoint increased by 1.1 M GdmCl compared to NI_3C and it is characterized by a $\Delta G_{(HX)}$ for the most slowly exchanging protons of approx. 12.5 kcal/mol (Figure 2.5(c)) which is 1.5 kcal/mol larger than NI_3C, demonstrating a higher overall stability of NI_3C_Mut5 throughout the internal repeats. Thus, in NI_3C_Mut5, the internal repeats display significantly higher protection than in NI_3C, although they are identical in sequence. Therefore, the redesign of the C-cap[7] has not only affected local unfolding events within the cap but greatly improved the coupling between the C-terminal cap and the rest of the protein. This improved packing of the C-cap against the internal repeats is also seen in the crystal structure of NI_3C_Mut5 (Kramer *et al.*, submitted). This improved coupling

of the C-cap is therefore transmitted across the whole protein and apparently also retards local fluctuations events (Figure 2.5 (c) see results from Ising model below).

2.3.3 Equilibrium denaturant unfolding of NI_3C and NI_3C_Mut5 analyzed by NMR

To obtain information on which regions unfold upon titration with GdmCl, we followed chemical shift changes at 20°C in [^{15}N,^{1}H]-HSQC spectra measured in the presence of increasing concentrations of GdmCl (0 to 7 M) with 200 and 300 µM ^{15}N-labeled NI_3C and NI_3C_Mut5, respectively. The buffer contained 50 mM phosphate, 150 mM NaCl at pH 7.4, as used previously in the CD- and fluorescence experiments.[8]

In the first steps from 0 to 2.1 M GdmCl several cross-peaks of the NI_3C wild-type C-cap (e.g. A141, Q142, T148-F150, G157, A162-Q166) completely disappear or move towards the random-coil region (8.0-8.6 ppm), while new peaks appear. Representative spectra are shown in Figure 2.7 and Figure S2.8. In contrast, peaks from residues of the internal repeats or from the N-cap are shifted much less, and their position can still be traced over that range of GdmCl concentrations. These data clearly indicate that the C-cap of NI_3C unfolds at denaturant concentrations lower than those required for global unfolding and are thus fully consistent with our earlier proposal of this unfolding pathway.[7] In contrast, the peaks of the stabilized C-cap in NI_3C_Mut5 do not disappear, but solely shift with increasing concentration of denaturant. Signal dispersion in the HSQC spectrum of NI_3C collapses at 3.6-4 M GdmCl (Figure S2.8), indicating the absence of proper side-chain packing accompanied by increased backbone flexibility. It should be noted that both collective chemical shift changes, i.e., the ones involving the C-cap at 0-2.1 M GdmCl and the ones involving the other repeats at 3.6-4 M GdmCl, occur at lower GdmCl concentrations than the changes observed in the CD signature (at 3-4 M GdmCl and 5.5-6 M GdmCl, respectively).[8] This suggests that a loss of tight side-chain packing precedes the loss of secondary structure.

A more detailed analysis of the ^{15}N chemical shift changes as a function of denaturant concentration for NI_3C_Mut5 reveals that the largest changes are observed in the long loops connecting the repeats. This is particularly true for Asp but also for Lys residues in those parts, indicating that these residues are the primary targets of interaction with the denaturant for NI_3C_Mut5 (see Figure S2.9). For protons additional large chemical shift changes are observed at the termini of the α-helices on the face opposite to the long loops, most likely due to a slight destabilization of the helices at the ends. A similar picture emerges from the data recorded for NI_3C, but the fact that shifts cannot be monitored over such a large concentration range and the additional occurrence from signals due to the intermediate complicate data analysis. Nevertheless, in both cases the observed changes are gradual, and the ^{15}N{^{1}H}-NOE data recorded in the presence of denaturant reveal that more extensive structural changes occur only at higher concentrations of denaturant (see Figure S2.10).

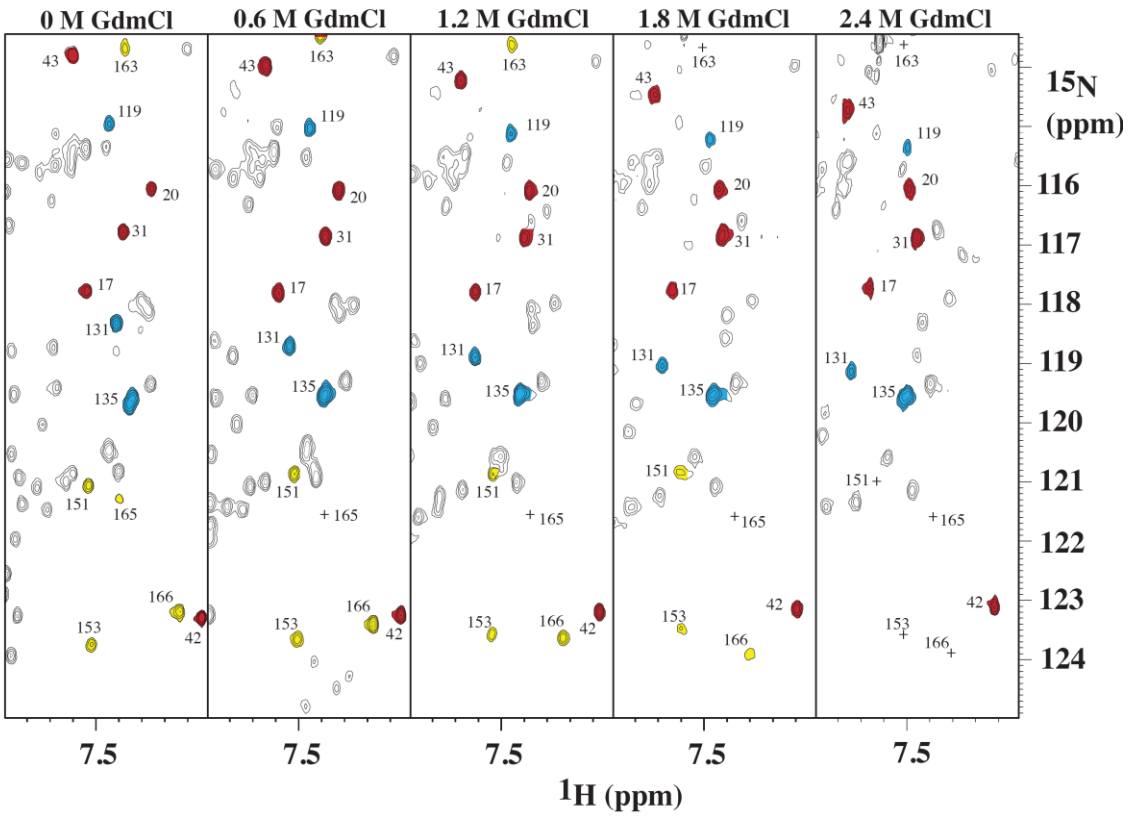


Figure 2.7 Expansion from 700 MHz [^{15}N,^{1}H]-HSQC spectra of NI^{3}C at GdmCl concentrations of 0, 0.6, 1.2, 1.8 and 2.4 M (from left to right) at pH 7.4 and 293K in the buffer used for the assignment. Peaks from the C-cap that disappear at elevated denaturant concentrations as well as peaks from internal repeats are annotated. Peaks from the C-cap are color-coded in yellow, from the N-cap in red and from the internal repeats in blue.

The ^{15}N{^{1}H}-NOEs were measured as a function of GdmCl to further investigate the destabilization of NI_3C and NI_3C_Mut5 by monitoring backbone dynamics of both proteins. At 4 M GdmCl the ^{15}N{^{1}H}-NOEs of NI_3C is approx. 0.25-0.3 for most residues, and the amide proton exchange is very fast (*vide infra*). These observations indicate that the structure is significantly destabilized (data not shown). Considering that proton chemical shifts are sensitive to side-chain conformations while the CD signal at 222 nm is mostly sensitive to secondary structure, and taking the information from the ^{15}N{^{1}H}-NOEs into account, we propose that the changes monitored by NMR (at 3.6-4 M GdmCl) characterize reduced packing of side-chains, and that the changes observed by CD (at 5.5-6 M GdmCl) reflect the complete loss of secondary structure.

The spectrum of NI_3C_Mut5 still displays reasonable signal dispersion up to 5 M GdmCl (Figure S2.11), while the transition identified from CD measurements is between 6.5 and 7.5 M GdmCl[7]. NI_3C_Mut5 in the absence of GdmCl is characterized by a ^{15}N{^{1}H}-NOE of 0.8, and the value is equally high both for residues of the helices as well as for residue of the long loops (Figure S2.12c). For GdmCl concentrations up to 4 M, values still generally remain higher than 0.6 (Figure S2.10), indicating that the structure remains intact and that motions are restricted to rather small fluctuations, although the decreased signal-to-noise in the spectra leads to a larger scatter of the data.

At 6 M GdmCl the $^{15}N\{^{1}H\}$-NOE adopts values about 0.25-0.3 (Figure S2.10). Importantly, the data indicate that tertiary structure in NI_3C_Mut5 is present at 4 M GdmCl, and this justifies measuring amide proton exchange at 4 M denaturant in order to probe for global unfolding events.
All data obtained so far have indicated that the redesigned C-cap of NI_3C_Mut5 has a major influence on the stability of the internal repeats as well. In order to get further insight into these stability differences, we measured hydrogen exchange in the presence of deuterated GdmCl to allow observation of global unfolding events. The data for NI_3C_Mut5 clearly reveal exchange in the most central repeat to be significantly slower than for the corresponding positions of the adjacent repeats. After $^{1}H/^{2}H$ exchange at 310K for approx. one month in the presence of 4 M GdmCl only peaks of residues from the most central repeat are visible in the [^{15}N,^{1}H]-HSQC spectra (Figure 2.8).

Measurements at 3.5 M in contrast to those at 4.0 M GdmCl allowed extraction of the exchange rate constants. These rates are depicted in Figure 2.9 and reveal that hydrogen exchange in GdmCl, in comparison to measurements in the absence of denaturant, is particularly accelerated in the I-1 repeat by up to 60-fold, whereas the corresponding changes for I-3 are usually smaller than 10-fold and those of I-2 are smaller still. As a result, the exchange is slowest in the central repeat. We like to point out that due to technical difficulties mainly arising from the high salt concentration the rates cannot be extracted with the desired precision (from duplicate experiments we estimate that errors can be up to 50% for some residues), but the relative trend is clear, and this is also obvious from the fact that only residues of the central repeat remain in spectra at later time points in 4 M GdmCl (Figure 2.8d).

In addition, we have also recorded exchange spectra at 333 K (60°C) in the absence of denaturant in order to promote global unfolding events (Figure 2.9b). Exchange for I-1 is accelerated up to twenty-fold, and below five-fold in I-2 and I-3. However, the differences between the repeats are small so that the central repeat is not exchanging more slowly.

Amide exchange was also measured for NI_3C at 1 and 2 M GdmCl. Measurements were complicated by the fact that many additional signals from the putative intermediate populate the spectra, leading to a decrease of signal intensity for many peaks and a large number of additional peaks that cannot be assigned. The exchange rates of the few assigned residues that can be determined reliably do apparently not differ amongst each other significantly, and they also do not differ from the slowly exchanging unassigned residues (see Figure S2.13).

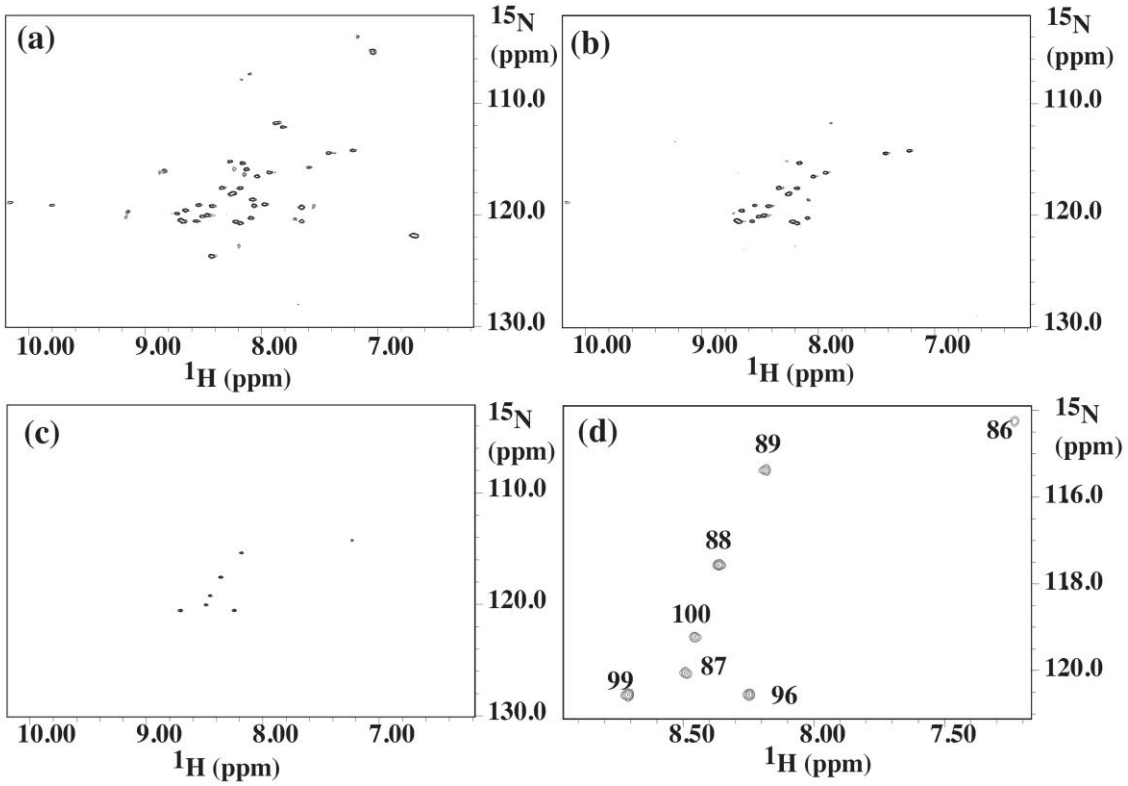


Figure 2.8 600 MHz [^{15}N,^{1}H]-HSQC spectra of NI$_3$C_Mut5 in presence of 4 M GdmCl, measured after dissolving in $^{2}H_2O$ after 0 hours (a), 58 h (b) and 670 hours (28 days; c and d), 310 K. The buffer used contained 50 mM phosphate, 150 mM NaCl, pH 7.4. Resonances remaining after 28 days are annotated with the corresponding residue number in the expansion shown in (d).

In addition, we tested pH-induced unfolding of NI$_3$C and NI$_3$C_Mut5. In these experiments the proteins were rebuffered to pH 3.5 and [^{15}N,^{1}H]-HSQC spectra were measured. In general, the intensity of signals outside the random-coil range decrease over hours to days at 310 K, accompanied by formation of new peaks, and the protein slowly precipitates. The ^{15}N{^{1}H}-NOEs mostly assume negative values indicating that the proteins unfold (data not shown). However, no quantitative data can be obtained.

2.3.4 **Comparison to calculations based on the Ising model**

An Ising model considers each repeat as an individual folding unit, characterized by an individual free energy ΔG, linearly dependent on denaturant, and stabilized by interactions with its folded neighbors, characterized by a coupling energy J [8,11]. The parameters of the Ising model describe the change in stability of the repeat as a function of environmental conditions and the magnitude of next-neighbor coupling to capture the behavior of the whole protein. In comparison to previous calculations,[8] in the present study the model was extended to take into account the difference in stability between the N- and C-caps by introducing different sets of stability parameters for the N-cap, for the original C-cap and for the mutated C-cap (see Materials & Methods).

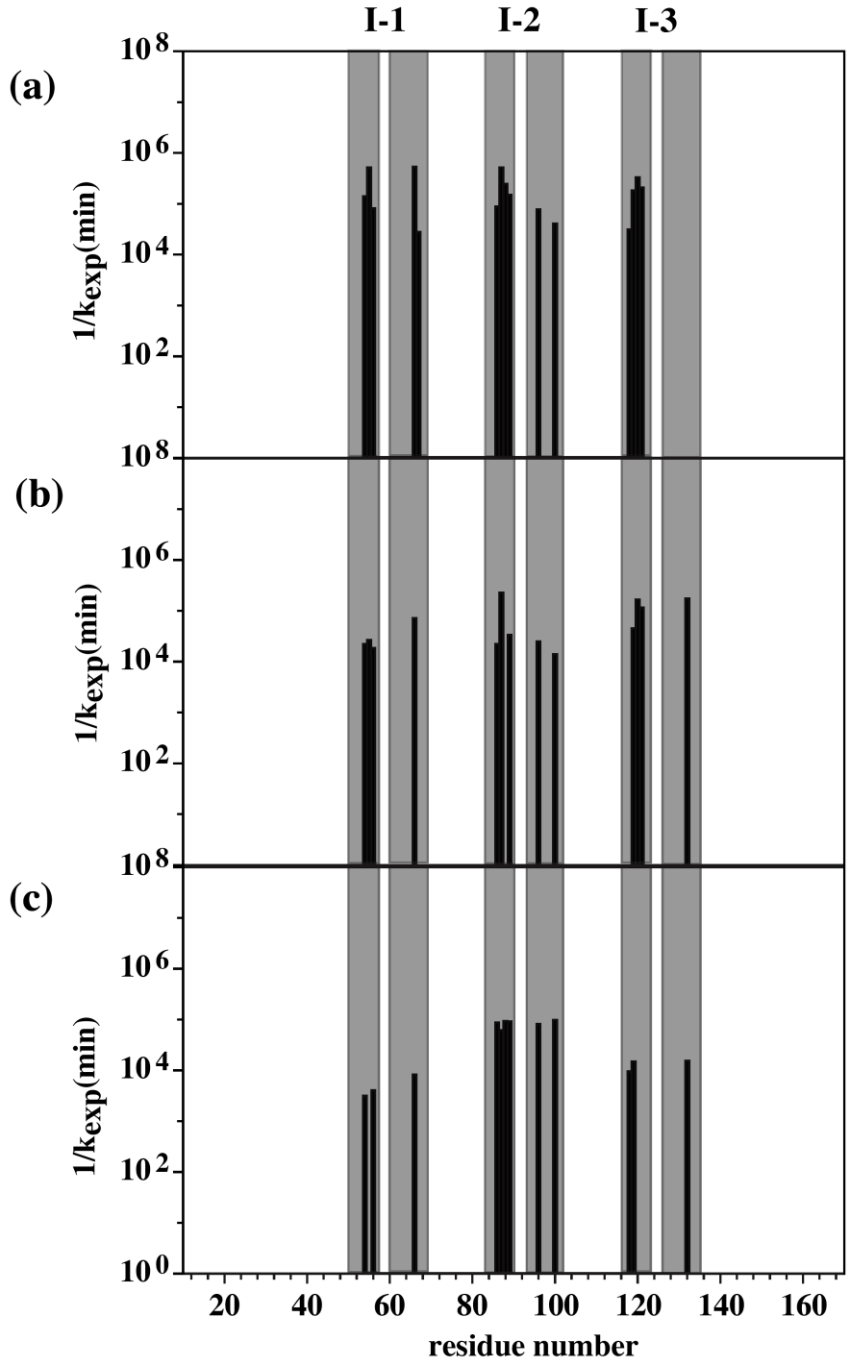


Figure 2.9 Protection factors for selected, slowly exchanging residues of the internal repeats I-1, I-2 and I-3 of NI_3C_Mut5 measured in native buffer at 310 K (a) or 333 K (b) or in 3.5 M GdmCl at 310 K (c).† The following conversion parameter have been used for the calculations of kch: log kA 1.5, log kB 9.66, log kW -1.28 (a,b), log kA 1.5, log kB 10.08 and log kW -1.825 (c)39,where kA,kB and kW are second-order rate constants for acid, base and water-catalyzed hydrogen exchange in poly-D,L-alanine.

† We noticed that the protection factors in the measurements at 333 K or 3.5 M GdmCl surprisingly increase for many residues when compared to the measurement in buffer at 310 K. We explain it by the fact that the temperature increase or the addition of denaturant does not significantly destabilize the protein in the most stably folded regions. As a result the only moderately accelerated exchange in those parts does not (over)compensate the large increase in the intrinsic exchange rate constant k_{ch}.

To avoid overfitting, the model was globally fit to a large set of data, comprising the CD-monitored equilibrium unfolding of five proteins[7, 8] (NI_1C, NI_2C, NI_3C, NI_1C_Mut5 and NI_3C_Mut5) and the PF values for the caps in the absence of denaturant reported in the present work.The PF values for the internal repeats have been excluded from the fit. Indeed, Ising model predictions obtained with the earlier version of the model show significantly larger protection factors for the internal repeats than the experimental values (data not shown). These reasons led us to conclude that protection factors of the internal repeats, despite involving extremely slow exchange rates, are seemingly the effect of local fluctuations in the native state. Exchange rates should be much slower if they were caused by complete, reversible unfolding of a single repeat, the cooperative unit of the Ising model.

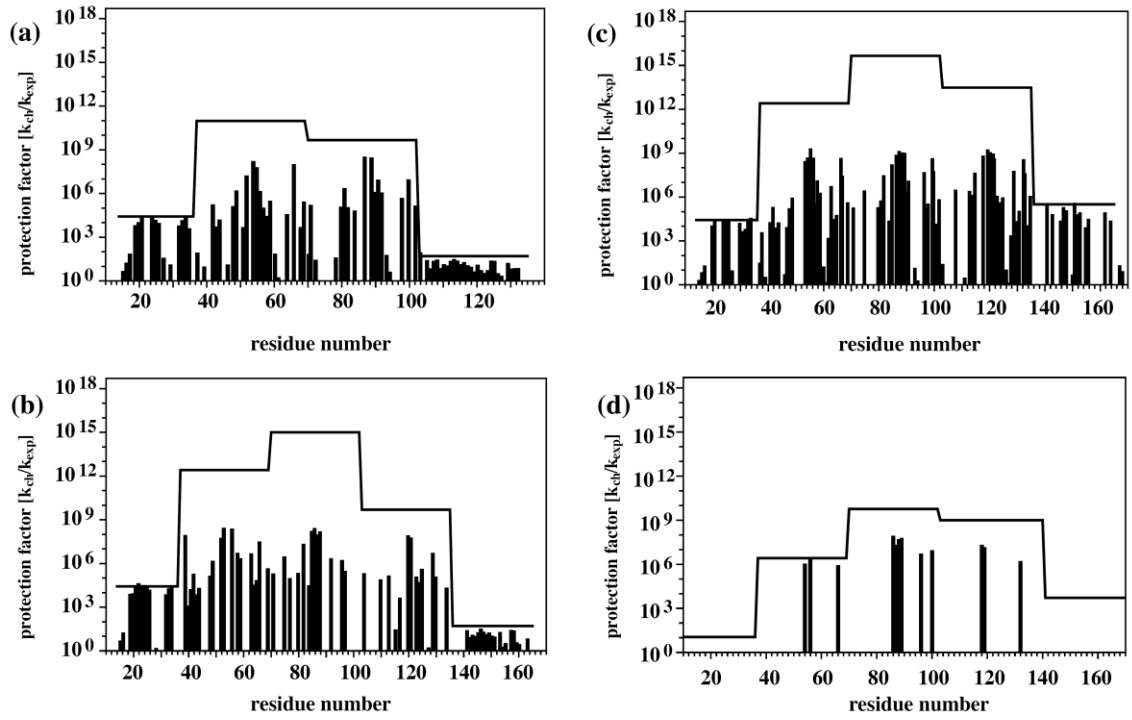


Figure 2.10 Expectations of the protection factorsat 293 K derived from an Ising-type folding model for NI_2C (a), NI_3C (b) and NI_3C_Mut5 (c) in absence of denaturant and NI_3C_Mut5 in presence of 3.5 M GdmCl (d). The experimental data derived from $^1H/^2H$ exchange are shown.

If the PF of the internal repeats are excluded for the reason of local fluctuations, the model can accommodate the experimental data (see Figure S2.14 for the CD-data and Figure 2.10 for the PF values of the caps). The extracted model parameters (see Table 1) for the repeat coupling (J), the stability of the internal repeats as well as for the average stability of the original caps are in good agreement with our previous computations (differences of about 1 kcal/mol). The parameters provide a quantitative measure of the difference in stability between capping repeats. In the absence of denaturant, the original C-cap is about 5 kcal/mol less stable than the mutant C-cap and almost 4 kcal/mol less stable than the N-cap.

2.4 Discussion

Repeat proteins allow testing some fundamental aspects of protein folding and stability, because of the ease with which the folded domain can be systematically increased in size. This knowledge also has practical consequences, as it will influence the design of libraries of repeat proteins for applications.[5, 18] For ankyrin repeat proteins, individual repeats are not structurally stable,[19] and the favorable contributions to the overall fold are entirely due to the interaction between neighboring repeats. It becomes thus of interest to study this effect quantitatively.

To simplify the analysis, we used full-consensus repeat proteins with identical sequence, having a different number of repeats, and with two different C-capping repeats. The original C-cap had previously been identified as the least stable repeat within the protein,[7] and its replacement by a designed C-cap stabilizes the whole protein. The challenge with full-consensus repeat proteins is that their resonances in NMR are significantly more difficult to assign due to the repetitive nature of the sequence. Nevertheless, we could demonstrate that full backbone assignments are possible using a suitable set of 3D triple-resonance experiments in combination with PRE data. This allowed us to use a variety of NMR experiments to characterize the denaturation behavior at residue resolution, such as $^{1}H/^{2}H$ exchange over many orders of magnitude and the measurement of chemical shift and heteronuclear NOE as a function of denaturant.

2.4.1 Stability-determining role of the C-cap

We confirmed by direct NMR measurements that the wild-type C-cap is significantly less stable than the designed Mut5 C-cap. This had previously been deduced from Molecular Dynamics (MD) calculations and indirectly from global unfolding experiments, and the design of the new C-cap was also inspired by MD calculations.[7] Here we not only confirmed the relatively low stability of the original C-cap (taken from a natural ankyrin[6]) but also observe that there is a minor population of a second conformation both at the C terminus in NI_2C and NI_3C, encompassing the last 5 residues of the protein in a largely unfolded state, with the flexibility being proven by the $^{15}N\{^{1}H\}$-NOE data. Since two sets of resonances are observed, the inter-conversion between the two forms must be slow on the NMR time scale (slower than 4 ms). This puts an upper limit on the inter-conversion rate, such that the lifetime of both conformations is significantly longer than required for simple helix-coil transitions that can be as fast as 100 ns, indicating that additional interactions must be formed during the transition.

The original C-cap has frequently shown higher B-factors in crystal structures, and it is this region, which is least well superimposed between different DARPin structures. In summary, it appears that this original C-cap, which was derived from a natural ankyrin, is not optimally packed against the central repeats. In contrast, the newly designed Mut5 C-cap does not display a second set of resonances. It also shows no evidence of selective early loss of C-cap cross-peaks in GdmCl titrations, its $^{1}H/^{2}H$ exchange rates are at least

as slow as those of the N-cap and the crystal structure is characterized by a better packing against the internal repeats (Kramer et al., submitted). We have now direct evidence for the energetic consequences of this coupling, since the native-state hydrogen exchange is indeed significantly slower in the internal repeats of NI_3C_Mut5 than in NI_3C (Figure 2.5).

2.4.2 **Coupling of adjacent repeats largely influences the folding energy landscape**

The increase in stability of repeat proteins with the number of repeats has been observed for ankyrin repeats, armadillo repeats and tetratricopeptide repeats.[11, 12] One of the relatively simple quantifications to explain these relations is an Ising-type model.[10-12] In this model, each individual repeat is treated as an autonomous unit that is only coupled to its neighbors. For successful *in vivo* folding special capping repeats with a hydrophilic surface are required, which are of lower intrinsic stability than the internal ones (this is true also for the newly designed Mut5 C-cap). The expectation from the Ising-type model is that the stability of repeats depends on their position in the protein: the most central ones would be expected to be more stable than those closer to the caps, because the central ones have a lower probability of having an unfolded neighbor. This effect is further enhanced by the lower intrinsic stability of the capping repeats. In particular, the central repeat of a NI_3C molecule should be the most stable.

The NMR methods used here have allowed us now to investigate the stability of the individual repeats directly. In native buffer, $^1H/^2H$ exchange rates were found to be very similar for corresponding positions in the different repeats, and the central internal repeat does not exhibit significantly higher protection factors. Expectations of protection factors using the Ising model parameterized here are depicted in Figure 2.10. These calculations predict that protection factors for the internal repeats are higher than those of the capping repeats, as is found experimentally. However, the PF of the internal repeats are lower than predicted from the model and do not show the expected highest PF for the central repeat. This is most probably due to local fluctuations, which can formally be seen as an exchange from the native state. The computed $\Delta G_{(HX)}$ values of NI_3C and NI_3C_Mut5 are thus smaller than those from the GdmCl-denaturation experiments measured by CD,[7, 8] and hence in native buffer, local fluctuations must significantly contribute to $^1H/^2H$ exchange. Thus, local fluctuations not only equalize the observed exchange rates of the internal repeats, but also make them faster than the values expected from repeat unfolding.

H/D exchange directly from the native state has been proposed for the consensus tetratricopeptide-repeat (CTPR)[15] to accommodate deviations from expectations of the Ising model. However, this is difficult to distinguish from local fluctuations. The differences between measured and H/D exchange rates expected from the Ising model in the full consensus DARPins are much greater than the differences found in CTPR, possibly due to the much higher overall stabilities of the DARPins, leading to an extremely rare complete unfolding of single repeats. There may be an intrinsic occurrence of small fluctuations that limits the maximal PF.

The discrepancy of ΔG values for global unfolding as determined from the CD measurements in presence of denaturant, and those computed from the measured exchange rates of the most slowly exchanging residues, thus indicates that these exchange events are not triggered by global unfolding. The protein NI_3 that lacks the C-cap displays a good quality [^{15}N,^{1}H]-HSQC spectrum with signal dispersion similar to NI_3C (Figure S2.15). However, the protein tends to precipitate over days at 310 K and its amide proton exchange is accelerated. Nevertheless, this well-dispersed spectrum demonstrates that species in which single repeats are unfolded are still stable for short periods of time, and may thus be intermediates in larger local unfolding and refolding events. Such subpopulations of NI_3C or NI_3C_Mut5 can either refold or continue unfolding. The higher intrinsic stability of the Mut5 C-cap increases its probability for refolding over that of the wild-type C-cap. In addition, the better coupling of the new C-cap to I-3 reduces local exchange in the internal repeats in NI_3C_Mut5 compared to NI_3C and propagates the stability increase through the rest of the molecule.

In the presence of high concentrations of GdmCl (e.g. 3.5-4.0 M) exchange rates of NI_3C_Mut5 are strongly accelerated, and differences in exchange rates now become clearly visible between the amide protons of the I-1 or I-3 and the I-2 repeats (Figure 2.10(d)). After approx. 4 weeks of exchange at 310 K in 4 M GdmCl only peaks from signals of the central repeat remain for NI_3C_Mut5 (Figure 2.8). At 3.5 M GdmCl rates of residues within I-2 are slower than for residues within I-1 and I-3, with the fastest rates being observed for I-1 (Figure 2.9). Conversely, at higher temperatures in the absence of denaturant, exchange is faster, but the difference between the internal repeats remain very small. This suggests that, when the whole protein is destabilized enough by denaturant, ^{1}H/^{2}H exchange by global unfolding of one repeat eventually becomes faster than exchange by local fluctuations, and the differences expected from the Ising-type model are indeed observed. In other words, the native DARPins are so stable and unfold so slowly in native buffer that local fluctuations put an upper limit to the residence time of a proton. However, under conditions of increased denaturant concentrations, partially unfolded species do not predominantly refold but rather continue unfolding such that under these conditions the exchange rates at least potentially report on unfolding of individual repeats.

The measurements of exchange in 3.5 M GdmCl are consistent with the stair-case appearance expected from the predictions based on the Ising model (cf. Figure 2.9(c), Figure 2.10(d)). We noticed, however, that the differences between exchange rates of I-1, I-2 and I-3 in the presence of denaturant are still smaller than those predicted based on an Ising model, which has been parameterized from optical spectroscopy to monitor unfolding and PF data for the capping repeats in buffer solution from the present work (Figure 2.10). This discrepancy is most likely due to the fact that even under destabilizing conditions local fluctuations still significantly contribute to ^{1}H/^{2}H exchange. However, in contrast to the measurements at 310 K in native buffer (Figure 2.9(a), local fluctuations do not completely dominate exchange so that differences in stability become experimentally accessible (Figure 2.9(c)).

Interestingly, the disappearance of signal dispersion in the [^{15}N,^{1}H]-HSQC spectra upon adding denaturant occurs well below the transition observed in the CD measurements.

We have interpreted the collapse of signal dispersion as due to local fluctuations resulting in reduced packing of side chains that, however, does not lead to overall removal of tertiary structure. In fact, it may be a hallmark of the electrostatic nature of the inter-repeat interactions that they persist even if side chain packing is loosened due to the less steep distance-dependence in comparison to other types of interactions (e.g. van der Waals interactions or hydrogen bonds).

The slow H/D exchange rates in Figure 2.9(c) and Figure 2.10(d) were collected at 310 K while the data used to fit the Ising model were measured at 293 K.[8] A possible temperature dependence of stability may thus introduce some additional changes, which are not taken into account by the procedure used here to translate exchange rates from 310 K to 293 K.[16] For example, a different dependence rate of stability loss of the N- and C-cap on temperature would also affect the adjacent repeats, whose stability depends on the adjacent cap being folded and could thus introduce a difference between repeats I-1 and I-3. Furthermore, at elevated temperature or in presence of GdmCl as denaturant the electrostatic interactions that mutually stabilize the neighboring repeats are reduced in strength. In either case differences in stability become visible because unfolding of entire repeats starts to contribute to the observed exchange.

Nevertheless, even under destabilizing conditions, local exchange therefore continues to contribute to an extent that mostly levels out the differences between the repeats.

The Ising models[10-12] that have been proposed represent a coupled system of independent 2-state folders, corresponding to the repeats. Whether the inherent assumption, namely that the individual repeats unfold themselves cooperatively, is valid or not cannot easily be judged based on our data. More complicated models, involving partially denatured states of individual repeats, where e.g. only one of the helices is unfolded allowing accelerated $^{1}H/^{2}H$ exchange, would certainly also be compatible with our measurements. Different architectures of repeat proteins may differ in this respect, due to the intrinsic stability of the repeats and the strength of coupling.

2.4.3 **Stabilizing features of the DARPins**

Our NMR analysis has now allowed pinpointing structural features in the DARPins that are particularly important for stability. This will be valuable for the future development of new variants and additional libraries.

Significantly reduced exchange rates for amide protons are observed for the central positions of the helices (Figure 2.6) in all proteins investigated in this work. In addition, selected positions within the long loops that connect the repeats are highly protected. Such high protection within loop regions is not necessarily expected. These protections highlight residues that are forming crucial interactions between repeats (*vide supra*)[7, 20] and thereby significantly contribute to the coupling and the overall stability of the proteins.

We observed that the exchange rates are slower for residues of helix 1 compared to those of helix 2 of the same repeat module. We attribute this to the additional interactions helix 1 experiences with residues of the preceding repeat-connecting loop (especially through the conserved histidines of the TPLH motif). Moreover, the

additional solvent-shield provided by these loop residues may slow down local unfolding events. Interestingly, this first helix is devoid of intrinsic helical propensity, as predicted by the program AGADIR[21] (Figure S2.16), underlining the enormous importance of inter-repeat interactions for the stability of these proteins. This also implies that once such stabilizing interactions of the first helix are broken the respective repeat is expected to unfold. Isolated modules have been shown to be largely unfolded.[19] Moderate helical propensities were computed only for the first helix of the N-cap, which has no preceding loop, and the second helix from each internal repeat.

Residues used for randomization in the DARPin library are mostly charged in the full-consensus DARPins, forming a tight network of attracting charges, involving loop residues as well as residues at the termini of the helices. This was postulated to generate additional interactions to make the full-consensus DARPins extraordinarily stable proteins.[7, 20] The stabilizing interactions are formed by Asp and Lys residues from the loops and Arg and Glu residues in helix 1. It is therefore particularly interesting to note that around the charged loop residues the largest ^{15}N chemical shift changes occur upon addition of GdmCl. Screening the favorable inter-repeat electrostatic interactions involving Asp and Lys residues in the loop may therefore be an important factor in denaturant-induced protein unfolding of these proteins. Nonetheless, GdmCl can safely be expected[22] to predominantly act on the main chain and nonpolar groups, and indeed NI_3C_Mut5 does not unfold in 6 M NaCl (data not shown).

The proposed importance of ionic interactions for the stability of these repeat proteins, especially the role of the histidines (*vide supra*), is further demonstrated by the fact the proteins unfold at room temperature at pH 3.5. Spectra recorded at that pH display reduced signal dispersion, and most residues are characterized by negative $^{15}N\{^1H\}$-NOEs.

2.4.4 **Comparison with other repeat proteins**

Native-state HX exchange studies have also been performed with natural repeat proteins, which are much less stable than the full-consensus DARPins. The tumor suppressor AR protein p16[23] displayed very fast exchange for most residues within the dead time of the HX experiment, and only six residues could be followed. Those rates are consistent with the $\Delta G°$ measured in urea unfolding equilibrium.[24] This demonstrates that the extraordinary stability of the DARPins is not due to a high intrinsic stability of the ankyrin fold *per se*, but rather a consequence of consensus engineering.

Partially folded species have also been characterized in two consensus tetratricopeptide-repeat (CTPR) proteins by an extensive $^1H/^2H$ exchange study.[15] In CTPR proteins, the outermost helices have much lower probability to be folded than the central helices, and these results are also consistent with an Ising model, requiring, however, the additional assumption that hydrogen exchange can take place also in the folded state of the protein. This may be equivalent to local fluctuations. As the slowest exchange rates are still 10-100 times faster than those measured here (PF about 10^4-10^7 for CTPR3[25] vs. 10^4-10^8 for NI_2C and NI_3C and 10^4-10^9 for NI_3C_Mut5), this suggests that local fluctuations are also a consequence of protein stability.

2.5 Conclusions

Hydrogen exchange data as well as the denaturant-induced unfolding studies conducted in this work indicate that the stability of the full-consensus ankyrin repeat proteins is greatly dependent on the coupling between repeats, most dramatically demonstrated by the stability enhancing cap mutations that are propagated throughout the whole protein. Denaturant-induced unfolding, followed by $^{1}H/^{2}H$ exchange, is consistent with an Ising-type description of equilibrium folding of NI_3C_Mut5, while native state exchange seems to be significantly governed by local fluctuations to allow exchange when unfolding events are too slow in these extremely stable proteins. Extraordinarily slowly exchanging protons indicate a stable core structure in the DARPins, which combines hydrophobic shielding with favorable electrostatic interactions. Key interactions between the loops and the helices cause a mutual stabilization.

2.6 Materials and Methods

2.6.1 Protein biochemistry and production of spin-labeled proteins

The repeat proteins NI_2C, NI_3C and NI_3C_Mut5 and the respective Cys mutants (see below) were expressed in the *E. coli* strain M15 (Qiagen) in M9 minimal medium containing ^{15}N-NH_4Cl as the sole nitrogen source: 5 ml of overnight culture (LB medium, 1% glucose, 100 mg/l ampicillin and 25 mg/l kanamycin, 37°C) was used to inoculate 1-l cultures (M9 medium, 1% glucose, 150 µM thiamine, 30 mg/ml ampicillin and 25 mg/ml kanamycin, 37°C). At OD_{600} = 0.6 (6 to 8 hours), the cultures were induced with 500 mM IPTG and further incubated for 4 hours.

For expression of triple ($^{15}N/^{13}C/^{2}H$) labeled proteins *E. coli* cells were adapted to high levels of $^{2}H_2O$ by streaking cells on agar plates with increasing content of $^{2}H_2O$ (LB agar in 50%, 100% $^{2}H_2O$) prior to inoculation. A 5 ml overnight culture (LB medium in $^{2}H_2O$, 1% ^{13}C-glucose, 100 mg/l ampicillin and 25 mg/l kanamycin, 37°C, 20 h) was used to inoculate 500 ml cultures (M9 medium in $^{2}H_2O$, 0.5% ^{13}C-glucose, 15 mg/l ampicillin and 13 mg/l kanamycin, 37°C). Cultures were induced at OD_{600} = 0.6 (after 12-18 hours) and further incubated for 4 hours. Purification was performed as described previously.[6, 8] To facilitate complete back-exchange of amide protons, an additional unfolding and refolding step was included in the purification procedure of the deuterated proteins. After loading the cell lysate, a Ni-NTA column was equilibrated with GdmCl-containing running buffer (5.5 M GdmCl for the back-exchange of NI_2C and 7 M GdmCl for the back-exchange of NI_3C and NI_3C_Mut5). Proteins were then incubated in the respective GdmCl solution on the column for 2 hours at 20°C. Stepwise refolding was subsequently achieved by lowering the GdmCl content in 0.5 M steps to 0 M GdmCl prior to elution of the proteins. The purity was checked by SDS-PAGE and the correct molecular mass verified by mass spectrometry ([^{15}N]-NI_2C (experimental 14.55 kDa, theoretical 14.56 kDa), [^{15}N]-NI_3C (exp. 18.12 kDa, theor. 18.13 kDa), [^{15}N]-NI_3C_Mut5 (exp. 18.39 kDa, theor. 18.40 kD), [^{15}N,^{13}C,^{2}H]-NI_2C (exp. 15.82 kDa, theor. 15.99 kDa), [^{15}N,^{13}C,^{2}H]-NI_3C (exp. 19.73 kDa, theor. 19.93 kDa), [^{15}N,^{13}C,^{2}H]-NI_3C_Mut5 (exp. 20.05 kDa, theor. 20.23 kDa)). In the case of the $^{15}N/^{13}C/^{2}H$ labeled proteins, we calculated a degree of deuteration of 78.9% (NI_2C), 80.3% (NI_3C) and 82.7% (NI_3C_Mut5). Yields for the ^{15}N-labeled proteins from 1 l were 90 mg for NI_2C, 50 mg for NI_3C and 91 mg for NI_3C_Mut5, and for the [^{15}N,^{13}C,^{2}H]-labeled proteins 16 mg for NI_2C, 26 mg for NI_3C and 10 mg for NI_3C_Mut5.

2.6.2 **Production of spin-labeled proteins:**

Cys residues for attaching spin-labels were introduced at positions 28, 150 and 155 using the following primers

```
D28Cfor:  5′-TTTTGGTCAGGACTGCGAAGTTCGTATCC-3′
D28Crev:  5′-TTTATACGAACTTCGCAGTCCTGACCAGCACG-3′
F150Cfor: 5′-TTTGTAAGACCGCTTGCGACATCTCCATCG-3′
F150Crev: 5′-TTTGATGGAGATGTCGCAAGCGGTCTTACCG-3′
D155Cfor: 5′-TTTGACTTAGCGATCTGCAACGGTAACGAGG-3′
D155Crev: 5′-TTTTCGTTACCGTTGCAGATCGCTAAGTCGAACGG-3′
```

The cysteine mutants were generated by inverse PCR using the respective forward and reverse oligonucleotide from the original expression plasmid pSW_NI_2C, pSW_NI_3C, and pSW_NI_3C_Mut5, with Pfu Turbo polymerase (1 min at 95°C; followed by 18 cycles of 30 sec at 95°C, 1 min at 55°C and 10 min 68°C; followed by 5 min at 68 °C, standard Pfu Turbo buffer). The remaining vector was digested with *Dpn*I for 3 h, purified, chemically competent *E. coli* XL1-Blue cells were transformed and the plasmids sequenced using standard techniques.

The cysteine mutants of NI_2C, NI_3C and NI_3C_Mut5 were incubated with 500 mM dithiothreitol (DTT) in PBS_{150} (50 mM phosphate, 150 mM NaCl, pH 7.4) for 30 min to ensure reduction of any intermolecular disulfide bridges. DTT was removed on a PD-10 column (Amersham) in PBS_{150} and the protein was immediately labeled at pH 6.4 under N_2 protection with (1-oxyl-2,2,5,5-tetramethyl-Δ3-pyrroline-3-methyl)-mercaptyl (MTSL, Toronto Research Chemicals) by adding MTSL spin-label (stock solution 100 mM in DMSO) to a final concentration 3.8 mM to 800 µM protein. The reaction was allowed to proceed for 4 hours at room temperature in the dark.[26, 27] Excess MTSL was removed on a PD-10 column in PBS_{150}, and the labeled protein was concentrated using amicon concentrator tubes. The success of the labeling reaction was verified by SDS-PAGE and mass spectrometry, where the peak from the spin-labeled protein complexes (SL) was the predominant one and unlabeled protein was a very minor species (< 5 %) ([^{15}N]-NI_2C-D28C-SL (exp. 14.72 kDa, theor. 14.73 kDa), ^{15}N-NI_3C-D28C-SL (exp. 18.29 kDa, theor. 18.30 kDa), ^{15}N-NI_3C-F150C-SL (exp. 18.26 kDa, theor. 18.27 kDa), ^{15}N-NI_3C_Mut5-D28C-SL (exp. 18.56 kDa, theor. 18.57 kDa), ^{15}N-NI_3C_Mut5-D155C-SL (exp. 18.56 kDa, theor. 18.57 kDa).

2.6.3 **Spin-label experiments:**

In order to identify residues in proximity to the spin-label, 2D ^{1}H-^{15}N-HSQC spectra of the Cys-mutants for both the spin-labeled and non spin-labeled species were recorded with 150-200 µM protein. Disulfide bond formation in the latter was suppressed by addition of 300 mM DTT. Cross-peaks were integrated in both spectra and their intensity ratio calculated.

In order to exclude intermolecular "bleaching" effects, the respective Cys mutant was expressed in non-labeled medium (LB) and coupled to MTSL. This species was mixed with ^{15}N-labeled Cys mutant protected with N-ethylmaleimide (NEM) to avoid dimer formation. For NEM coupling the cysteine mutants of NI_2C, NI_3C and NI_3C_Mut5 were incubated in 500 mM DTT in PBS_{150} (50 mM phosphate, 150 mM NaCl, pH 7.4) for 30 min to ensure disruption of all disulfide bridges. DTT was removed on a PD-10 column (Amersham) in PBS_{150} and the protein was immediately labeled with NEM (Fluka) in PBS_{150} at pH 7 under N_2 protection. A tenfold molar excess of freshly prepared NEM in water was added, and the reaction was allowed to proceed for 2 hours at room temperature. Excess NEM was removed on a PD-10 column in PBS_{150}, and the labeled protein was concentrated. Successful NEM-labeling was verified by SDS-PAGE and mass spectrometry (NI_3C-D28C-NEM (exp. 18.23 kDa, theor. 18.24 kDa), ^{15}N-NI_3C-F150C-NEM (exp. 18.20 kDa, theor. 18.21 kDa), ^{15}N-NI_3C_Mut5-D28C-NEM (exp. 18.50 kDa, theor. 18.51 kDa), ^{15}N-NI_3C_Mut5-D155C-NEM (exp. 18.50 kDa, theor. 18.51 kDa). Both proteins were mixed in a 1:2 (^{15}N-NEM: ^{14}N-SL) ratio. The ratio of peak integrals from spectra of this mixture and a reference sample of ^{15}N-NEM without ^{14}N-SL at the same protein concentration was evaluated.

2.6.4 **NMR Spectroscopy and Data Evaluation**

NMR experiments were recorded using 700 µM solutions of NI_2C, NI_3C or NI_3C_Mut5 in 50 mM phosphate buffer, 150 mM NaCl, pH 7.4. All NMR experiments were recorded at 310 K on Bruker Avance 600 or 700 MHz spectrometers equipped with cryoprobes. To avoid severe loss of sensitivity and prohibitively long pulse durations, samples with high content of GdmCl were measured in 3 mm NMR tubes. For backbone assignment an approximately 79% deuterated, uniformly [^{13}C,^{15}N]-labeled sample was used. All NMR experiments used pulsed field gradients, sensitivity-enhancement schemes and water suppression through coherence-selection.[28, 29] Deuterium decoupling was applied during relevant ^{15}N or ^{13}C evolution periods or delays. Experiments were selected from the Bruker standard pulse sequence library (with the sole exception of the MEXICO experiment). HNCACB / HN(CO)CACB spectra[30] and HN(COCACB)CG / HN(CACB)CG spectra[31] were used to link sequential amide groups via matching pairs of Cα/Cβ and Cγ resonances. In addition, spin systems were linked via common carbonyl resonances using HNCO and HN(CA)CO experiments.[32] The usage of the HN(CACO)NH[33] experiment that provided direct correlations of amide groups with sequential nitrogen resonances proved to be particularly useful during assignment of the proton-nitrogen correlation map. A proton-detected version of the steady-state ^{15}N{^{1}H} heteronuclear nuclear Overhauser effect sequence was used for measurement of the heteronuclear NOE.[34]

All spectra were processed in TOPSPIN using mirror-image linear prediction for constant-time evolution periods. Spectra were mainly analyzed in the program CARA,[35] while batch processing of exchange spectra was accomplished with the programs XEASY[36] and SPSCAN.

2.6.5 **Measurement of amide proton exchange**

Slow amide proton exchange was measured by deuterium exchange experiments conducted at both 310 and 290 K. The ^{15}N-labeled proteins were lyophilized from the native buffer. To start the ^{1}H/^{2}H exchange reaction, the proteins were dissolved in 99.9 % $^{2}H_2O$ to yield concentrations of 750 µM, and a first HSQC spectrum was acquired 5 min thereafter. Further 28 time points were taken over 24 hours at 290 K. In a second set of experiments deuterium exchange was monitored over a period of 5 to 12 months at 310 K. These slow proton-deuterium exchange rates were measured using standard [^{15}N,^{1}H]-HSQC experiments. Finally, ^{1}H,^{2}H exchange in native buffer at 333 K (60°C) was measured for NI_3C_Mut5 by incubating the sample at 333 K in a water bath over the time span of 28 days, interrupted by short 1.5 hour measurements at 310 K (at this temperature to facilitate peak identification). Since exchange in native buffer is extremely slow at 310 K for residues of the internal repeats, the introduced error in the kinetics by measuring at 310 K instead of 333 K is negligible.

Faster exchange rates were derived from a series of MEXICO experiments.[13] The pulse sequence used contained doubly matched ^{13}C and ^{15}N filters. Moreover, to avoid the devastating effects of radiation damping on cryoprobes at 700 MHz a weak gradient was continuously applied during the recovery delay.

Hydrogen-deuterium exchange rates in the presence of 1 or 2 M GdmCl (NI_3C) or 3.5 and 4.0 M GdmCl (NI_3C_Mut5) were measured at 310 K by dissolving the lyophilized protein in fully deuterated GdmCl buffer of the corresponding GdmCl concentration. Up to 20 time points were taken over a period of 6 to 8 weeks.

For data collected at 290 K, the assignment (obtained at 310 K) was transferred to 290 K by a series of 2D [^{15}N,^{1}H]-HSQC spectra recorded in 2 K steps between 290 and 310 K.

2.6.6 **Data analysis of amide proton exchange**

Mono-exponential functions were fitted to the peak volumes of signals in the proton-deuterium exchange experiments using the Marquart-Levenberg algorithm:

$$I(t) = I_0 \cdot \mathrm{e}^{-k_{exp} \cdot t} + I_{\mathrm{inf}} \qquad (1)$$

Similarly, the following function was fitted to the peak volumes from the MEXICO experiments

$$\frac{I}{I_{ref}} = \frac{k}{-R_w + R_1 + k} \cdot (e^{-R_w t} - e^{-(R_1 + k)t}) \qquad (2)$$

where R_W denotes the T_1 of water (determined in a separate experiment), R_1 is the rate constant for the decay of longitudinal amide proton magnetization and k the rate constant for the exchange process of interest. Note that this function is slightly different from the one suggested by Hofmann *et al.*[37] to take into account the different treatment of water magnetization during the recovery delay necessary when recording data on cryoprobes.

Protection factors (PF) were calculated as the ratio k_{ch} / k_{exp}, where k_{exp} is the measured exchange rate and k_{ch} is the exchange rate of the amide in the unfolded state

based on peptide models (dependent on temperature, pH and neighboring amino acids).[9] The stabilizing free energy of the protecting structure can then be calculated as

$$\Delta G_{(HX)} = -RT \ln(k_{ex} / k_{ch}) \tag{3}$$

Exchange rates determined at 290 K were converted to the expected values at 310 K according to

$$k_{ex}(T) = k_{ex}(290)\,\mathrm{e}^{-\frac{E_a}{R}\left(\frac{1}{T}-\frac{1}{290}\right)} \tag{4}$$

where E_a = 19 kcal/mol and R = 1.986 cal/mol·K is the gas constant,[16] and the data recorded at 333 K were treated in an analogous fashion.

2.6.7 **Measurement of GdmCl or pH-induced equilibrium unfolding by NMR**

For measuring HSQC spectra in GdmCl the samples were equilibrated at the corresponding GdmCl concentration overnight at 20°C. The final GdmCl concentrations were determined by the refractive index. Because of the high salt content and signal broadening, spectra were recorded with increasing numbers of scans with increasing GdmCl concentrations. The assignment was transferred step-wise from 0 M to 3.6 M (NI_3C) and 5 M (NI_3C_Mut5) denaturant. The differences in ^1H and ^{15}N chemical shifts between 0 and 3.6 or 5 M denaturant for selected residues were plotted against the protein sequence.

For measurements of the proteins at pH 3.5 a different buffer (7.5 mM phosphate, 7.5 mM citric acid and 7.5 mM boric acid, 85 mM KCl) was prepared. The pH was quickly lowered from 7.4 to 3.5 by adding 20 ml of this pH 3.5-buffer to 50 µl the protein in PBS_{150} buffer. The solution was concentrated by centrifugation and re-diluted 4 times with this buffer in order to completely exchange the buffer. The first measurement was started approx. 2 hours after the initial mixing of buffers.

2.6.8 **Fit to Ising model**

The Ising model presented in Ref. 8 was extended to take into account the difference in stability observed between the N-cap and the original as well as the mutated C-caps by introducing separate sets of stability parameters for each of them, in addition to the set describing the internal repeats. The functional form for the effective conformational free energy is:

$$\Delta G_{Conf}\left(\{s_i\}; D, L, y\right) = \Delta G^I(D)\sum_{i=2}^{L-1} s_i + \Delta G^N(D)\cdot s_1 + \Delta G^y(D)\cdot s_L + J\sum_{i=1}^{L-1} s_i s_{i+1} \tag{5}$$

where s_i is a variable describing the state of the i-th repeat (1 when folded, 0 when unfolded), L is the total number of repeats in the protein, D is the denaturant concentration, and y indicates the type of C-cap of the protein (C0 for the original C-cap, or C1 for the mutated C-cap). The $\Delta G^x(D) = \Delta G_0^x + m^x \cdot D$ are the stabilities of the repeat of type x (with x=I, N, C0, C1 for the internal, N-cap, original C-cap and mutated

C-cap repeat, respectively) and depend linearly on the denaturant concentration. Thus, the model is characterized by 9 free parameters, i.e., the coupling parameter *J* and 4x2 stability parameters (ΔG_0^I, ΔG_0^N, ΔG_0^{C0}, ΔG_0^{C1}, m^I, m^N, m^{C0} and m^{C1}) instead of the previously used 5 parameter in ref. 8. To avoid overfitting, the model was globally fitted to a large set of data, comprising the CD-monitored equilibrium unfolding data[7, 8] and the PF values for the caps in the absence of denaturant reported in the present work. The PF values for the internal repeats have been excluded from the fit, as they are affected by local fluctuations. The fit to the CD data was obtained following the procedure described in reference 8. The fit of the PF data was obtained by assuming that the largest PF of repeat *i* is inversely proportional to the probability of observing the repeat unfolded, p_u,[15]

$$PF(D,i,L) \approx \frac{1}{p_u(D,i,L)} \tag{6}$$

with

$$p_u(D,i,L) = 1 - \sum_{\{s_j\}} P_L(\{s_j\};D) \cdot s_i \tag{7}$$

where $P_L(\{s_j\};D)$ is the probability of observing the conformation $\{s_j\}$ (see eq. 19 in ref. 8) and the sum is extended to all possible conformations. The global fit was obtained by minimizing the χ^2, which had a contribution from the PF and one from the CD data, weighted so that both sets of data contribute similarly. χ^2 was minimized using the Levenberg-Marquardt procedure and the errors reported on the parameters are standard deviations obtained from the inverse of the approximate Hessian matrix of the χ^2.
All amide proton, ^{15}N, Cα and Cβ chemical shifts of NI_2C, NI_3C and NI_3C_Mut5 have been deposited in the BMRB database under accession codes 16718, 16717 and 16716, respectively.

2.7 Acknowledgements

The authors thank Dr. Annemarie Honegger for software predicting the distance dependence of spin labels and Dr. Jiri Mares for writing scripts for the calculation of rate constants. This work was supported by SNF grant No CRSI00_122686 (to OZ) and the NCCR Structural Biology (to AP).

2.8 References:

1. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134**, 117-131.
2. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.* **293**, 151-160.
3. Kobe, B. & Kajava, A. V. (2000). When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.* **25**, 509-515.
4. Sedgwick, S. G. & Smerdon, S. J. (1999). The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem. Sci.* **24**, 311-316.
5. Binz, H. K., Amstutz, P., Kohl, A., Stumpp, M. T., Briand, C., Forrer, P., Grütter, M. G. & Plückthun, A. (2004). High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol.* **22**, 575-582.
6. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Plückthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332**, 489-503.
7. Interlandi, G., Wetzel, S. K., Settanni, G., Plückthun, A. & Caflisch, A. (2008). Characterization and further stabilization of designed ankyrin repeat proteins by combining molecular dynamics simulations and experiments. *J. Mol. Biol.* **375**, 837-854.
8. Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K. & Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *J. Mol. Biol.* **376**, 241-257.
9. Krishna, M. M., Hoang, L., Lin, Y. & Englander, S. W. (2004). Hydrogen exchange methods to study protein folding. *Methods* **34**, 51-64.
10. Zimm, B. H. & Bragg, J. K. (1959). Theory of the phase transition between helix and random coil polypeptide chains. *J. Chem. Phys.* **31**, 526-535.
11. Mello, C. C. & Barrick, D. (2004). An experimentally determined protein folding energy landscape. *Proc. Natl. Acad. Sci. U S A* **101**, 14102-14107.
12. Kajander, T., Cortajarena, A. L., Main, E. R., Mochrie, S. G. & Regan, L. (2005). A new folding paradigm for repeat proteins. *J. Am. Chem. Soc.* **127**, 10188-10190.
13. Gemmecker, G., Jahnke, W. & Kessler, H. (1993). Measurements of fast proton-exchange rates in isotopically labeled compounds. *J. Am. Chem. Soc.* **115**, 11620-11621.
14. Palmer, A. G. (2004). NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* **104**, 3623-3640.
15. Cortajarena, A. L., Mochrie, S. G. & Regan, L. (2008). Mapping the energy landscape of repeat proteins using NMR-detected hydrogen exchange. *J. Mol. Biol.* **379**, 617-626.

16. Bai, Y., Milne, J. S., Mayne, L. & Englander, S. W. (1993). Primary structure effects on peptide group hydrogen exchange. *Proteins* **17**, 75-86.
17. Kohl, A., Binz, H. K., Forrer, P., Stumpp, M. T., Plückthun, A. & Grütter, M. G. (2003). Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proc. Natl. Acad. Sci. U S A* **100**, 1700-1705.
18. Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O., Caflisch, A. & Plückthun, A. (2008). Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *J. Mol. Biol.* **376**, 1282-1304.
19. Mosavi, L. K., Minor, D. L. J. & Peng, Z. Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl. Acad. Sci. U S A* **99**, 16029-16034.
20. Merz, T., Wetzel, S. K., Firbank, S., Plückthun, A., Grütter, M. G. & Mittl, P. R. (2008). Stabilizing Ionic Interactions in a Full-Consensus Ankyrin Repeat Protein. *J. Mol. Biol.* **376**, 232-240.
21. Munoz, V. & Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* **1**, 399-409.
22. Pace, C. N. & Shaw, K. L. (2000). Linear extrapolation method of analyzing solvent denaturation curves. *Proteins* **Suppl 4**, 1-7.
23. Tang, K. S., Fersht, A. R. & Itzhaki, L. S. (2003). Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure* **11**, 67-73.
24. Boice, J. A. & Fairman, R. (1996). Structural characterization of the tumor suppressor p16, an ankyrin-like repeat protein. *Protein Sci.* **5**, 1776-1784.
25. Main, E. R., Stott, K., Jackson, S. E. & Regan, L. (2005). Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proc. Natl. Acad. Sci. U S A* **102**, 5721-5726.
26. Lietzow, M. A., Jamin, M., Dyson, H. J. & Wright, P. E. (2002). Mapping long-range contacts in a highly unfolded protein. *J. Mol. Biol.* **322**, 655-662.
27. Hornung, T., Volkov, O. A., Zaida, T. M., Delannoy, S., Wise, J. G. & Vogel, P. D. (2008). Structure of the cytosolic part of the subunit b-dimer of Escherichia coli F0F1-ATP synthase. *Biophys. J.* **94**, 5053-5064.
28. Kay, L. E., Keifer, P. & Saarien, T. (1992). Pure absorption gradient enhanced heteronuclear single-quantum correlation sepctroscopy with improved sensitivity. *J. Am. Chem. Soc.* **114**, 10663-10665.
29. Keeler, J., Clowes, R. T., Davis, A. L. & Laue, E. D. (1994). Pulsed-field gradients: theory and practice. *Methods Enzymol.* **239**, 145-207.
30. Shan, X., Gardner, K. H., Muhandiram, D. R., Rao, N. S., Arrowsmith, C. H. & Kay, L. E. (1996). Assignment of N-15, C-13(alpha), C-13(beta), and HN resonances in an N-15, C-13, H-2 labeled 64 kDa trp repressor-operator complex using triple-resonance NMR spectroscopy and H-2-decoupling. *J. Am. Chem. Soc.* **118**, 6570-6579.
31. McCallum, S. A., Hitchens, T. K. & Rule, G. S. (1998). Unambiguous correlations of backbone amide and aliphatic gamma resonances in deuterated proteins. *J. Magn. Reson.* **134**, 350-354.

32. Yamazaki, T., Lee, W., Arrowsmith, C. H., Muhandiram, D. R. & Kay, L. E. (1994). A suite of triple-resonance NMR experiments for the backbone assignment of N-15,C-13,H2- labeled proteins with high sensitivity. *J. Am. Chem. Soc.* **116**, 11655-11666.
33. Weisemann, R., Rüterjans, H. & Bermel, W. (1993). 3D triple-resonance NMR techniques for the sequential assignment of NH and ^{15}N resonances in ^{15}N- and ^{13}C-labelled proteins. *J. Biomol. NMR* **3**, 113-120.
34. Noggle, J. H. & Schirmer, R. E. (1971) *The nuclear Overhauser effect - chemical applications* (Academic Press, New York).
35. Keller, R. (2004) *The computer aided resonance assignment* (CANTINA Verlag, Goldau).
36. Bartels, C., Xia, T. H., Billeter, M., Güntert, P. & Wüthrich, K. (1995). The program XEASY for computer-supported spectral analysis of biological macromolecules. *J. Biomol. NMR* **6**, 1-10.
37. Hofmann, H., Weininger, U., Low, C., Golbik, R. P., Balbach, J. & Ulbrich-Hofmann, R. (2009). Fast amide proton exchange reveals close relation between native-state dynamics and unfolding kinetics. *J. Am. Chem. Soc.* **131**, 140-146.
38. Bai, Y., Milne, J. S., Mayne, L. & Englander, S. W. (1993). Primary structure effects on peptide group hydrogen exchange. *Proteins* **17**, 75-86.
39. Loftus, D., Gbenle, G.O., Kim, P.S. & Baldwin, R.L. (1986). Effects of denaturants on amide proton exchange rates: a test for structure in protein fragments and folding intermediates. *Biochemistry* **25**, 1428-1436.

2.9 Supplementary Material for Wetzel *et al.*

2.9.1 Some remarks considering the assignment:

For NI_2C, 125 (and thus 6 more than the maximally expected 119) cross-peaks were detected in the [^{15}N,^{1}H]-HSQC spectrum, for NI_3C 161 (151 expected) and for NI_3C_Mut5 152 (153 expected). In general, peaks due to residues from the His-tag and up to the first ankyrin residue (Asp13) were invisible at pH 7.4. Because of overlapping or missing peaks, we were not able to assign amide moieties of D13 in all three proteins and K101 in NI3C. The overall completeness of the backbone assignment was 99% for NI_2C, NI_3C and NI_3C_Mut5 (see alsoFigure S2.2, Figure S2.3, and Figure S2.4). Amide proton, 15N, Cα and Cβ chemical shifts of NI_2C, NI_3C and NI_3C_Mut5 have been deposited in the BMRB database under accession codes 16718, 16717 and 16716, respectively.

We have stated the importance that individual cross peaks are observed in the [^{15}N,^{1}H]-HSQC spectrum of the repeat protein. Resolving peaks along the amide proton frequency is in principle possible in ^{15}N-resolved NOESY experiments, but the resolution in the proton (F1) domain was often insufficient, and hence we either used the HN(CACO)NH experiment to exploit resolution in the ^{15}N domain or resorted to triple-resonance spectra with peak matching in ^{13}C. Nevertheless, the combination of knowledge of amide proton and nitrogen frequencies of sequential amide moieties greatly facilitated assignments. Since carbonyl chemical shifts usually display good signal dispersion, the 3D HNCO and HN(CA)CO experiments were recorded and additionally utilized for assignments.

2.9.2 **Supplementary Figures**

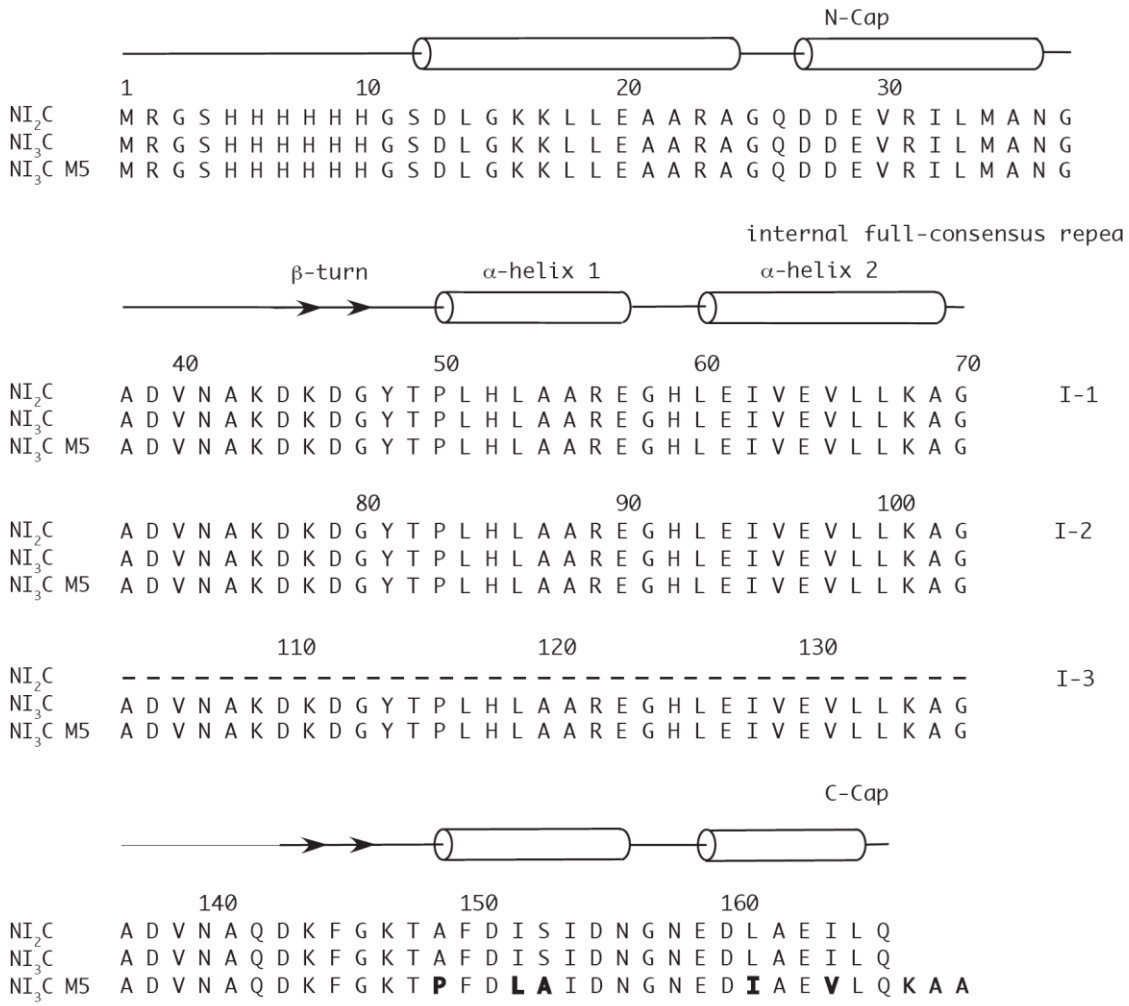


Figure S2.1 Sequence alignment of the full-consensus ankyrin repeat proteins NI_2C, NI_3C and the C-cap mutant NI_3C_Mut 5. The mutated residues in the C-cap are in boldface. Elements of secondary structure have been annotated according to the crystal structure of NI_3C (PDB entry 2QYJ).

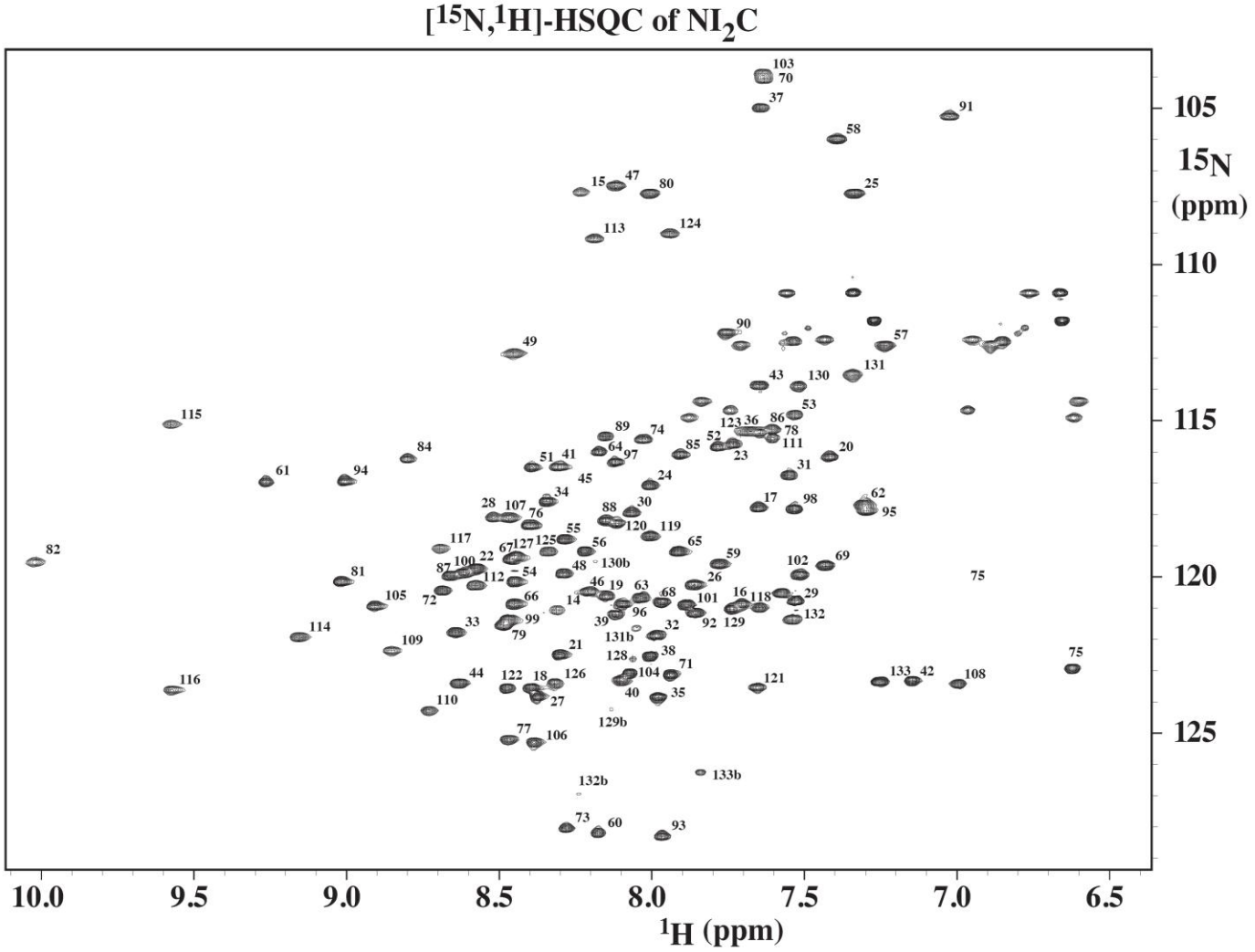


Figure S2.2 600 MHz [^{15}N,^{1}H]-HSQC spectrum of NI_2C, 310 K, in 50 mM phosphate, 150 mM NaCl, pH 7.4

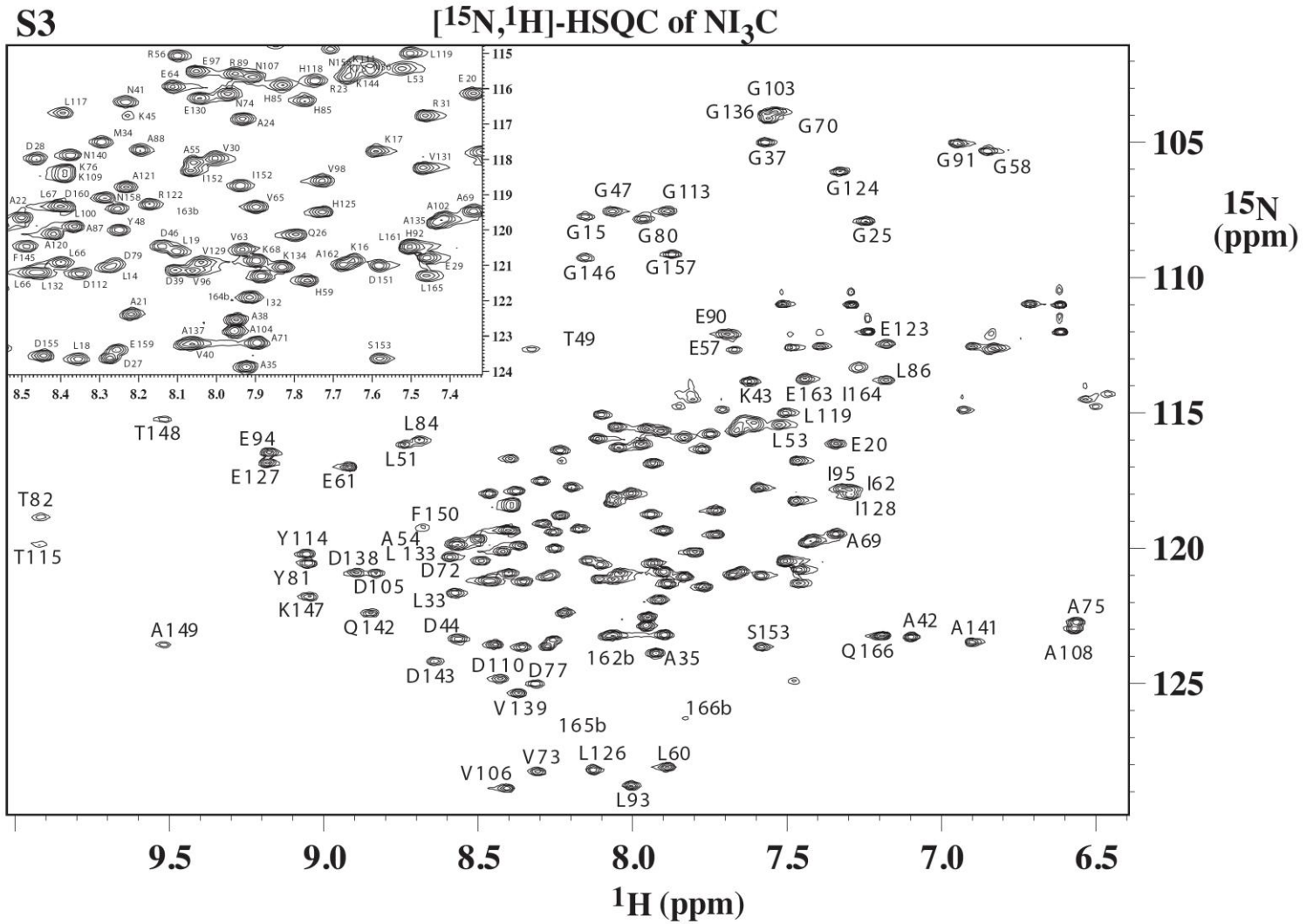


Figure S2.3 600 MHz [^{15}N,^{1}H]-HSQC spectrum of NI$_3$C, 310 K, in 50 mM phosphate, 150 mM NaCl, pH 7.4

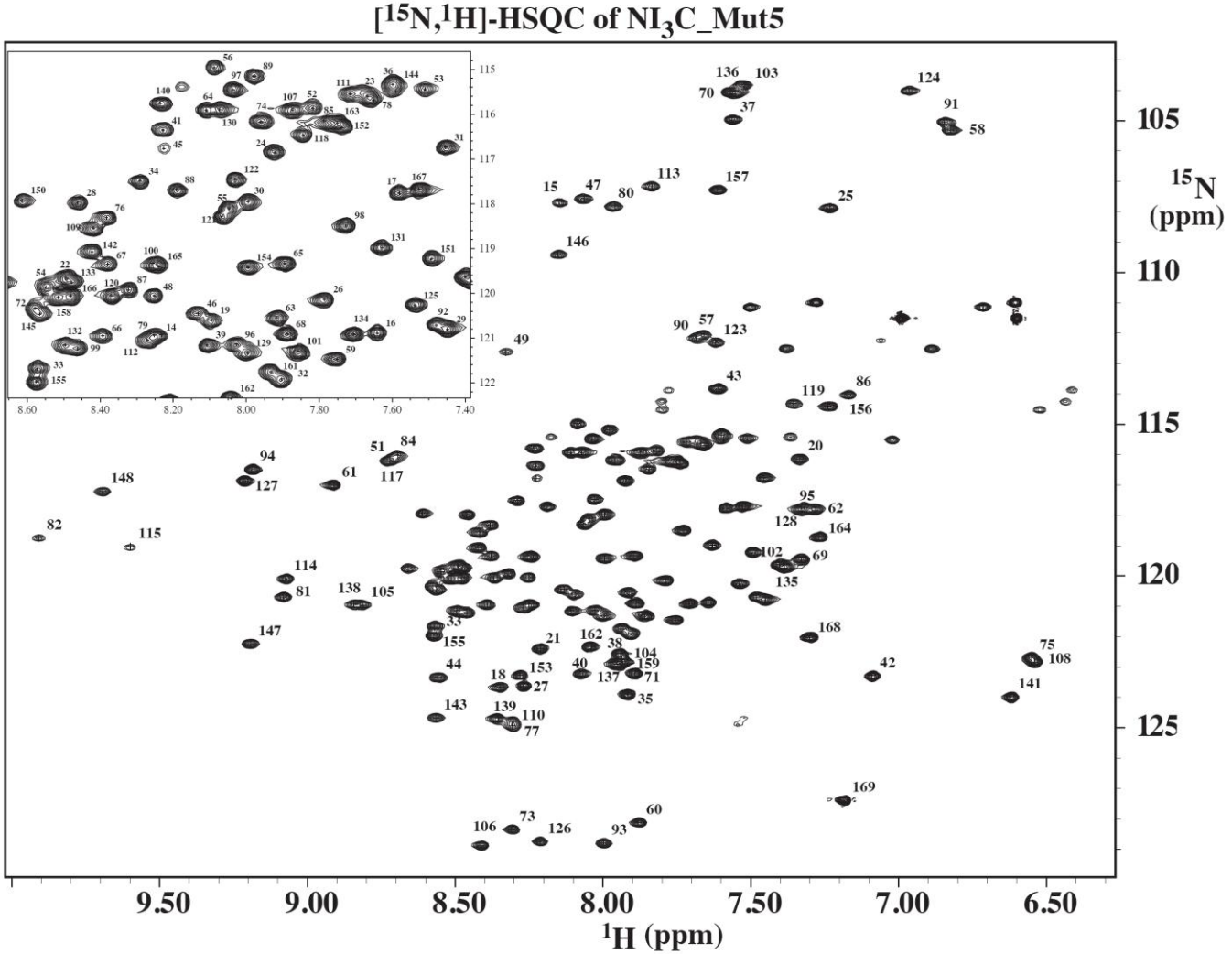


Figure S2.4 600 MHz [^{15}N,^{1}H]-HSQC spectrum of NI$_3$C_Mut5, 310 K, in 50 mM phosphate, 150 mM NaCl, pH 7.4

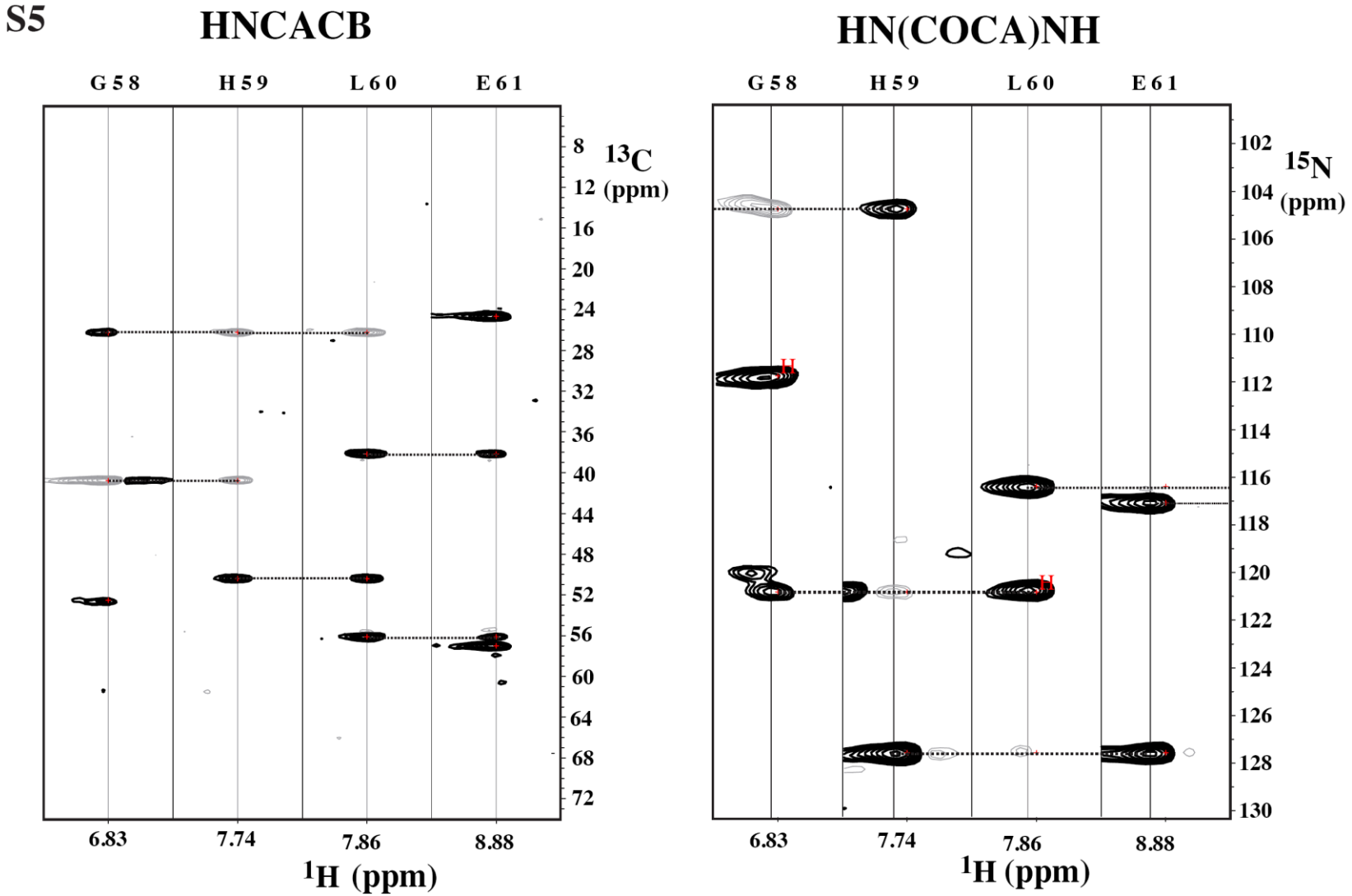


Figure S2.5 Representative strips from the 3D HNCACB (left) and HN(COCA)NH (right) spectra of NI_3C_Mut5

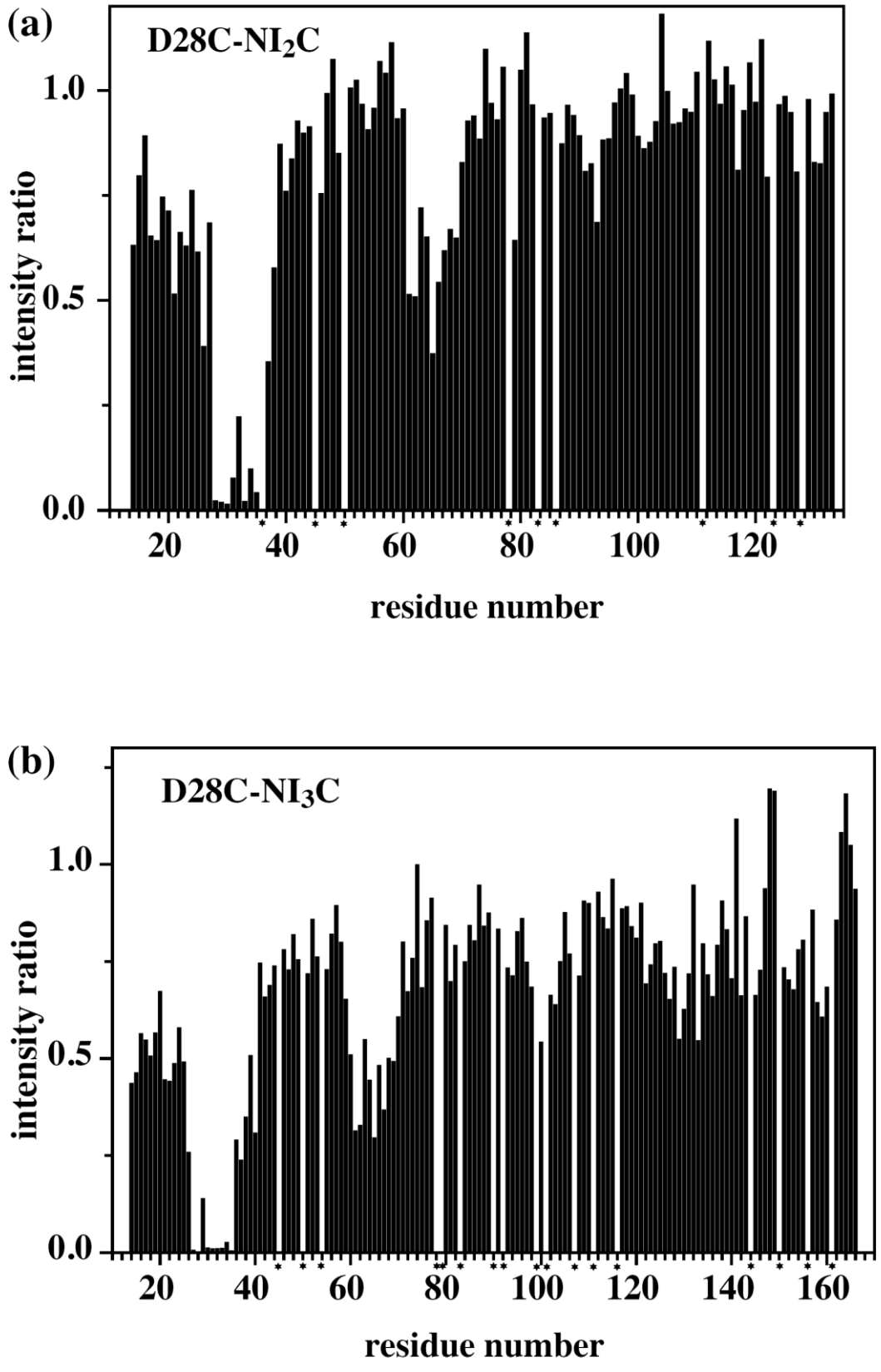


Figure S2.6 ***Intramolecular*** **attenuations of signal intensities of cross peaks in the 600 MHz [^{15}N,^{1}H]-HSQC spectra of MTSL-derivatized D28C-NI_2C (top) and D28C-NI_3C (bottom) relative to the non-spin labeled protein. Residues for which peaks could not be integrated reliably have been omitted and are marked by an asterisk directly below the horizontal axis**

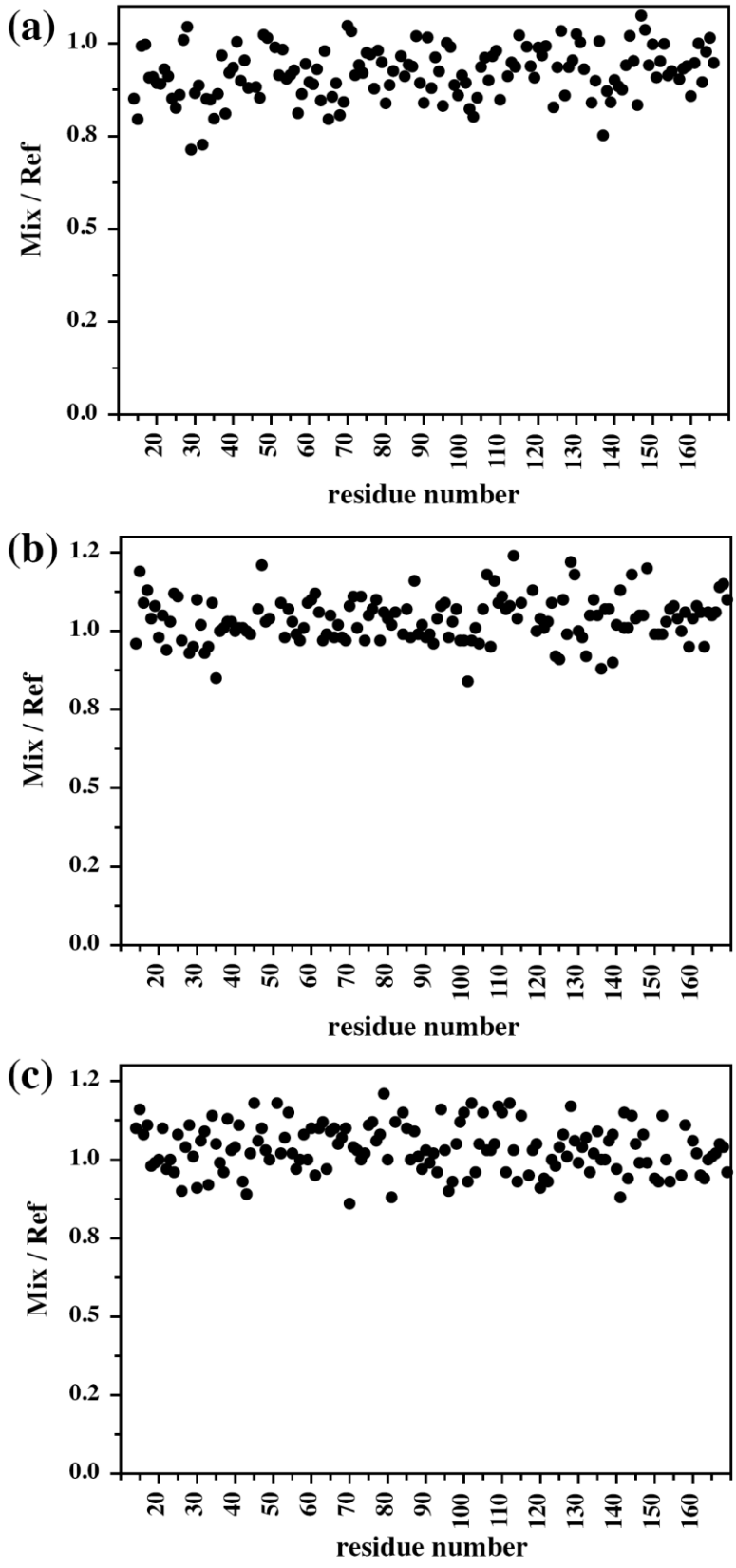


Figure S2.7 ***Intermolecular*** **attenuations of signal intensities of cross peaks in the 600 MHz [^{15}N,^{1}H]-HSQC spectra of a mixture of MTSL-derivatized unlabelled D28C-NI$_3$C (a), D28C-NI$_3$C_Mut5 (b) or D155C- NI$_3$C_Mut5 (c) with the non-spin labeled ^{15}N uniformly labeled corresponding proteins protected wit NEM. The ratio of signal in the mixed sample (*Mix*) over the 15N labeled sample alone (*Ref*) is plotted**

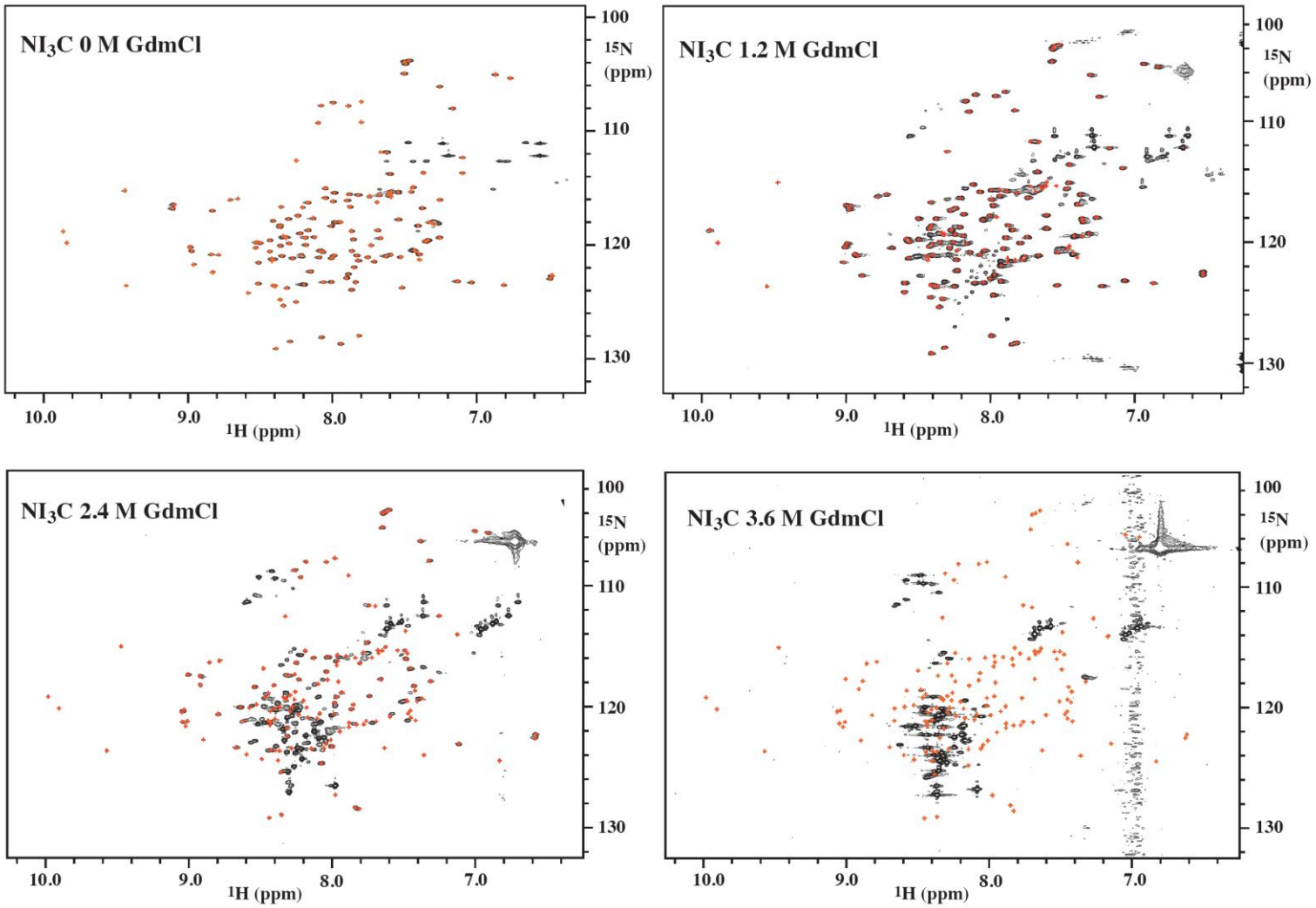


Figure S2.8 700 MHz [^{15}N,^{1}H]-HSQC spectra of NI_3C at 0 M (top left), 1.2 M (top right), 2.4 M (bottom left) and 3.6 M GdmCl (bottom right), 50 mM phosphate, 150 mM NaCl, pH 7.4, 310 K. Red crosses denote the positions of the original peaks in absence of GdmCl

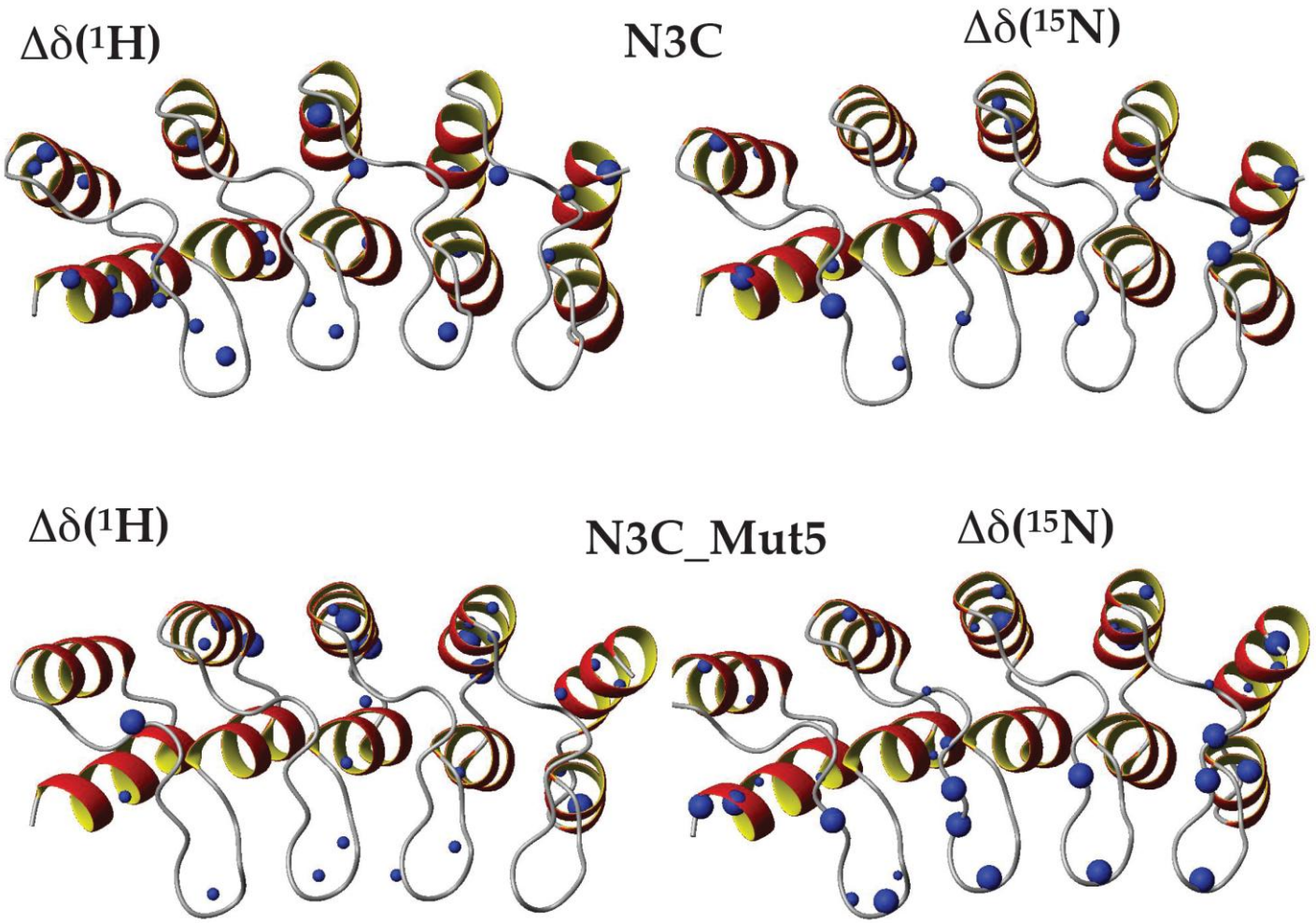


Figure S2.9 ^{1}H (left) and ^{15}N (right) chemical shift changes mapped onto the structure of NI_3C (top) and NI_3C_Mut5 (bottom). The size of the spheres corresponds to the absolute change in frequency between 0 and 2.1 M or 0 and 5.0 M GdmCl for NI_3C and NI_3C_Mut5, respectively

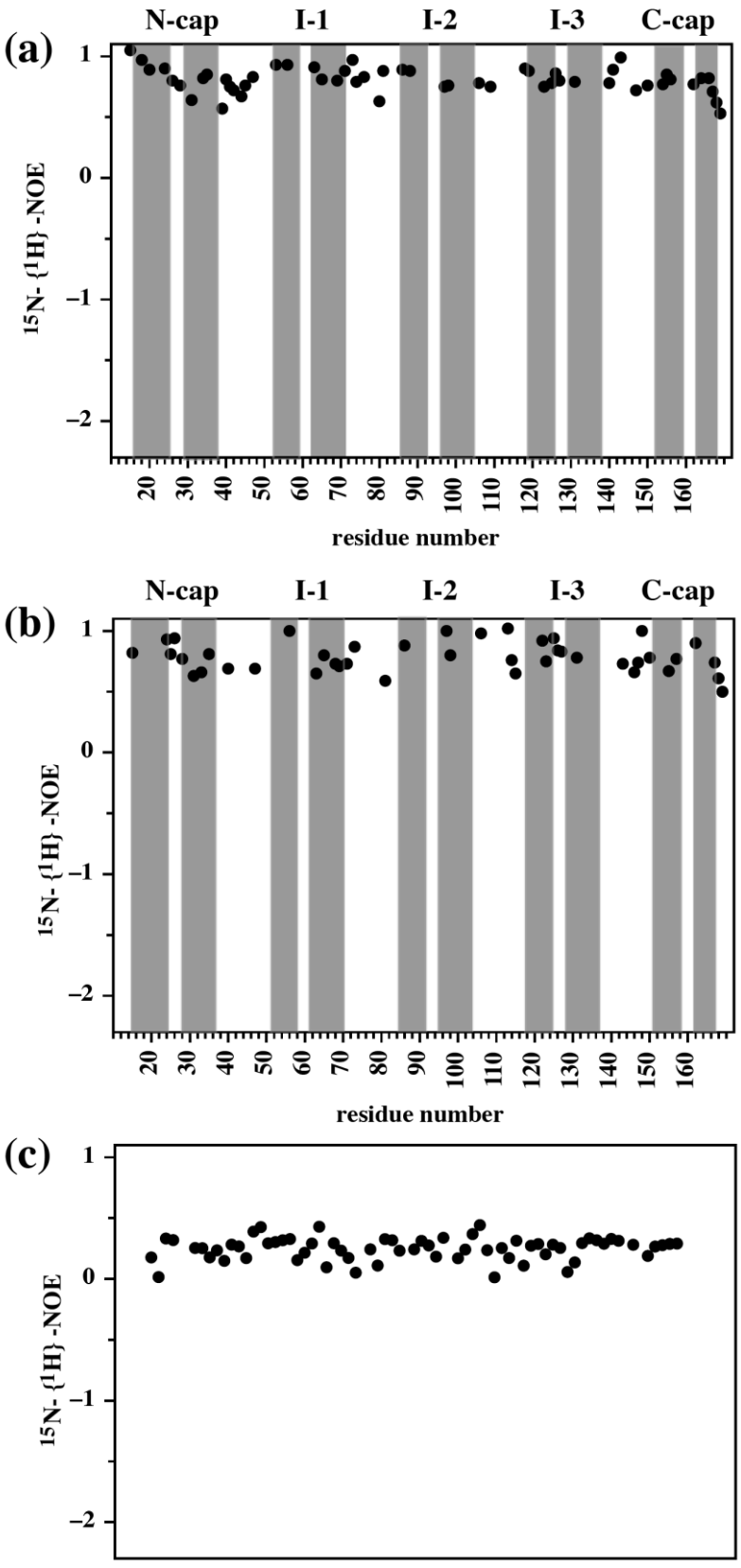


Figure S2.10 ^{15}N-{^{1}H}-NOE data of NI_3C_Mut5 in presence of 2 M (a), 4 M (b) or 6 M GdmCl (c) recorded at 600 MHz. In case of the data at 6 M GdmCl no assignments were available and the values on the x-axis are chosen arbitrarily

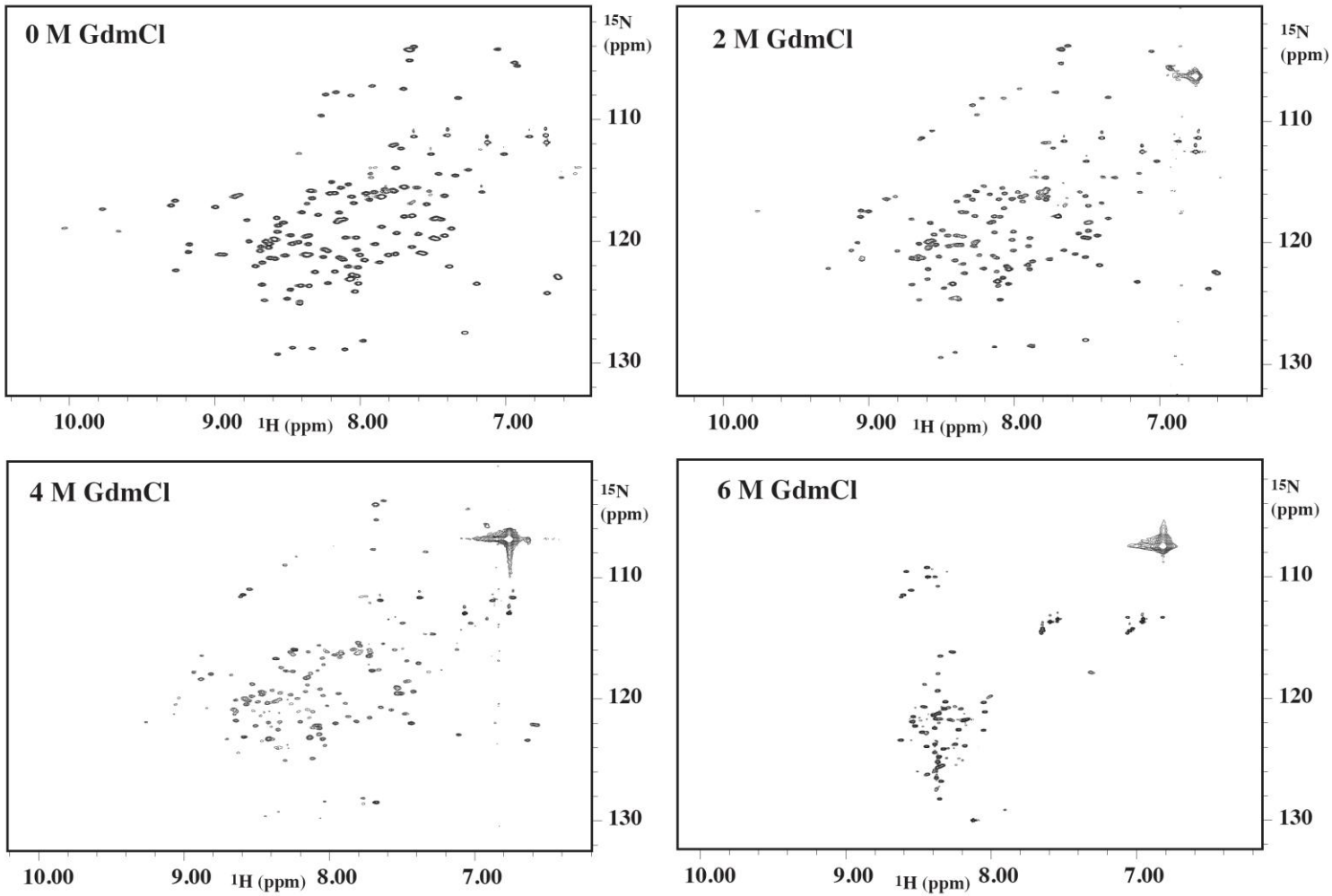


Figure S2.11 700 MHz [^{15}N,^{1}H]-HSQC spectra of NI$_3$C_Mut5 at 0 M (top left), 2.0 M (top right), 4.0 (bottom left) and 6.0 M GdmCl (bottom right), 50 mM phosphate, 150 mM NaCl, pH 7.4, 310 K

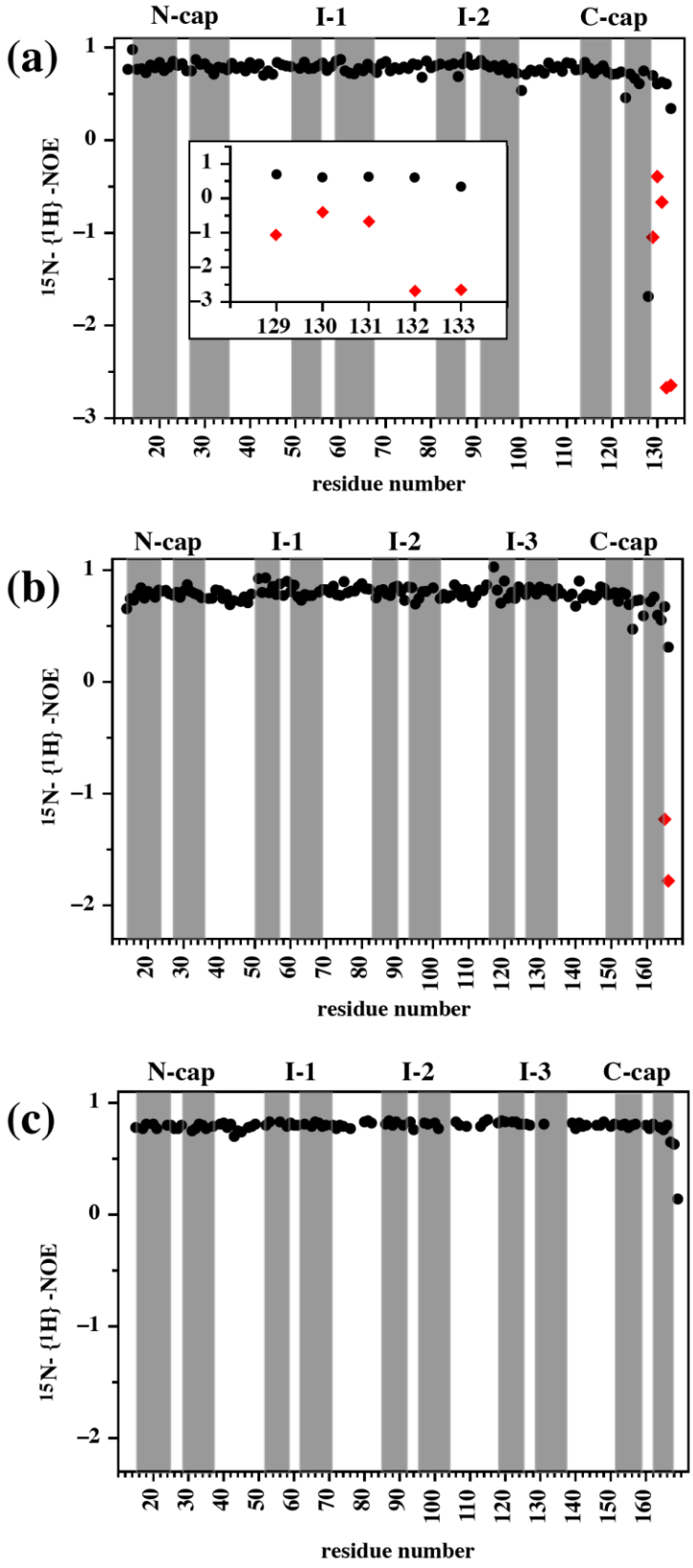


Figure S2.12 ^{15}N-{^{1}H}-NOE data of NI_2C (a), NI_3C (b) and NI_3C_Mut5 (c) recorded at 600 MHz on 0.7 mM solutions of the proteins in 50 mM phosphate, 150 mM NaCl, pH 7.4, 310 K. Gray background indicates the location of helices 1 and 2 in the repeats. Values for the second conformation of NI_2C and NI_3C are shown as red diamonds

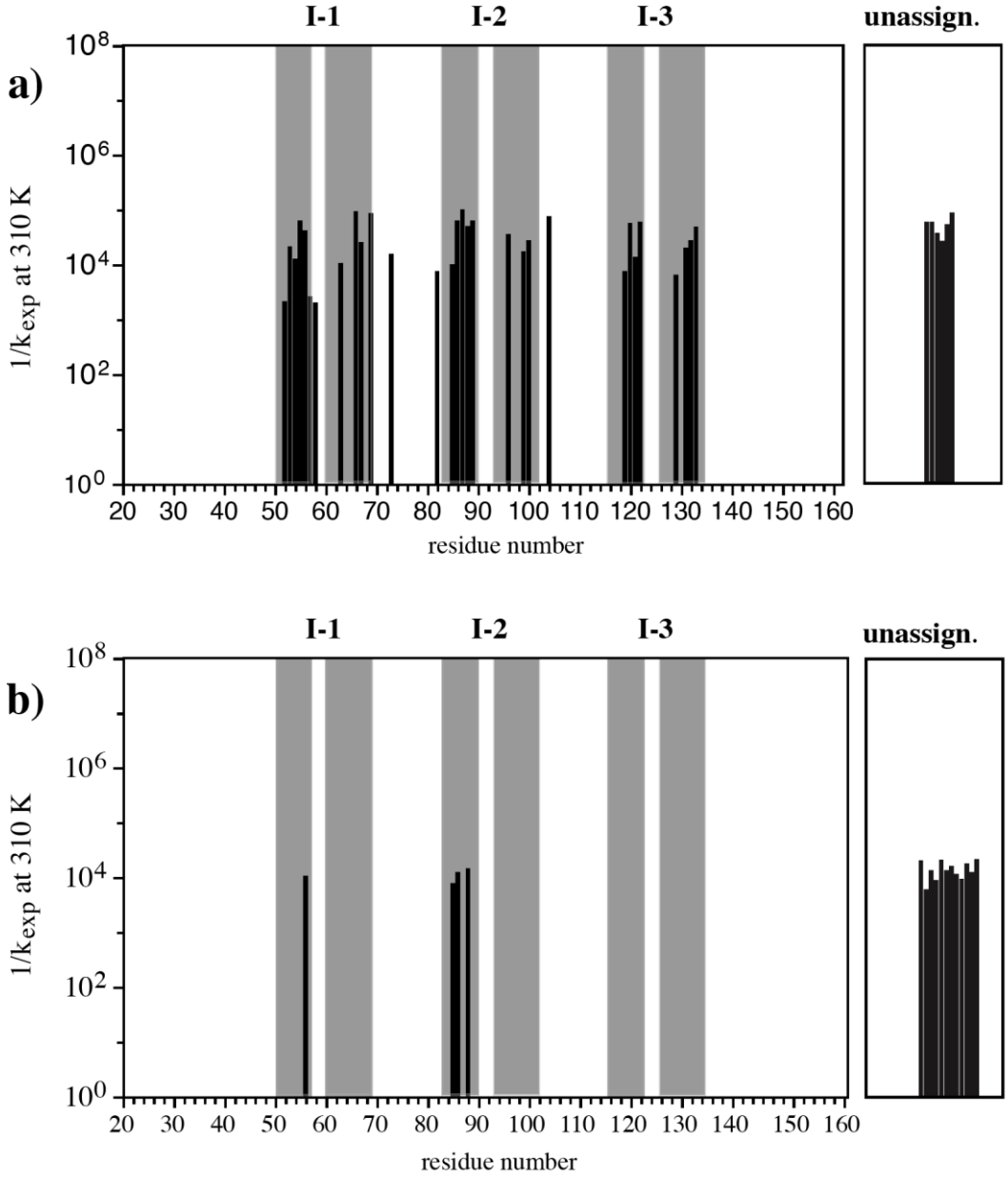


Figure S2.13 NI_3C in presence of 1 M (a) and 2 M (b) GdmCl. Values for residues that cannot be assigned are depicted in the separate panel on the right

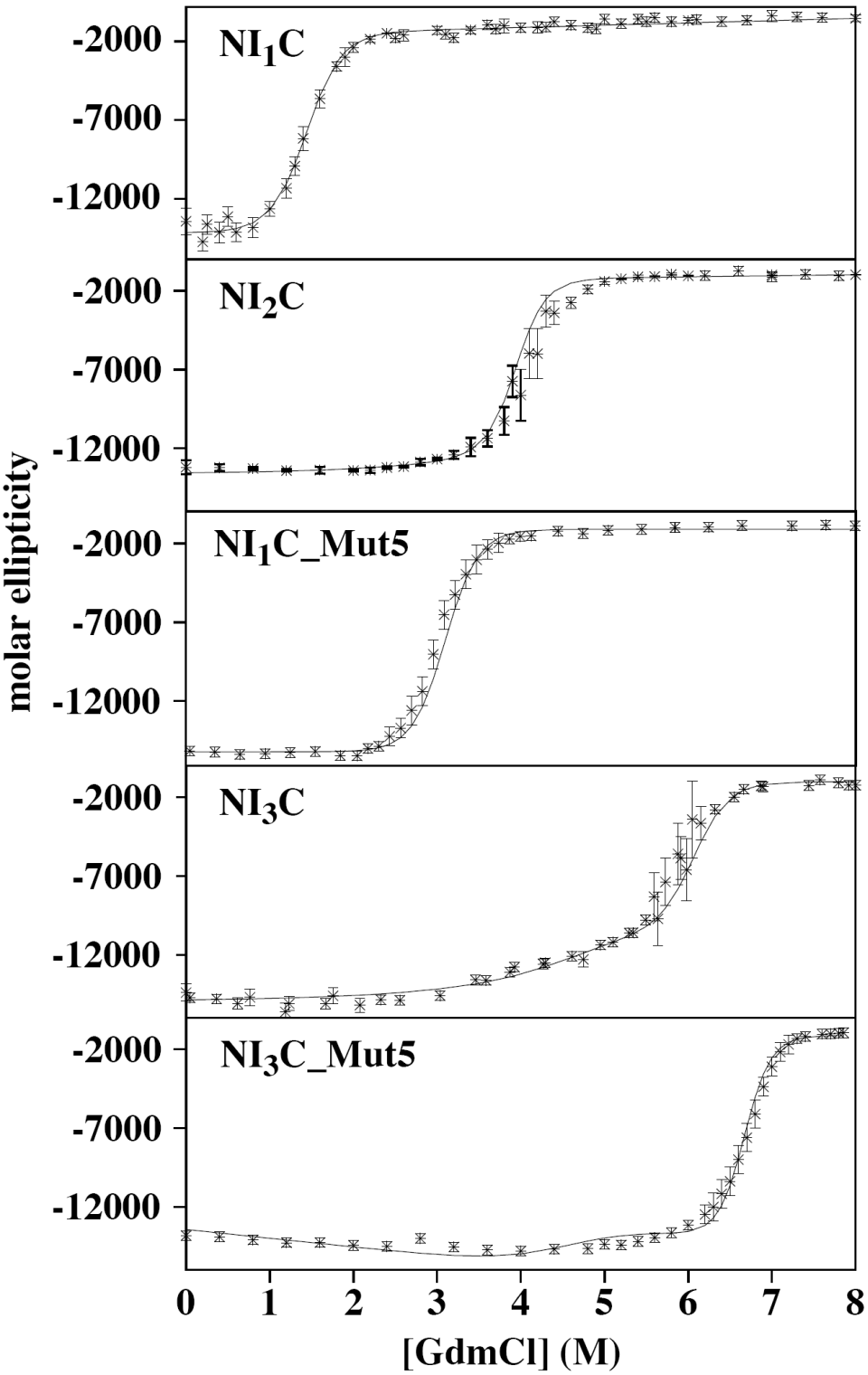


Figure S2.14 CD-monitored denaturation curves of NI_1C, NI_2C, NI_3C, NI_1C_Mut5 and NI_3C_Mut5 (asterisks with error-bars) and the Ising model fit (solid lines) obtained using equation 5 and the parameters reported in Table 2.1

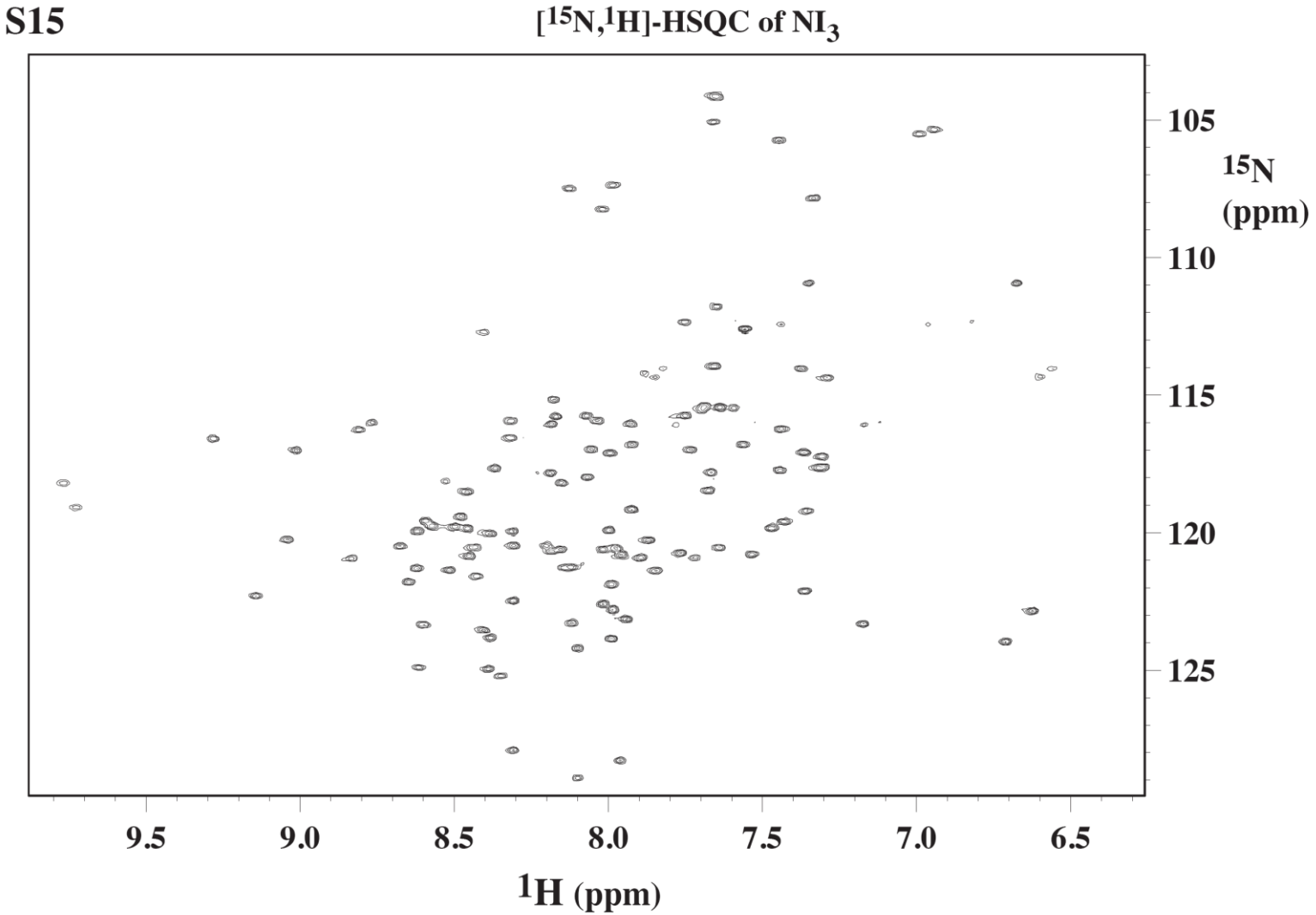


Figure S2.15 600 MHz [^{15}N,^{1}H]-HSQC spectrum of NI_3, 310 K, in 50 mM phosphate, 150 mM NaCl, pH 7.4

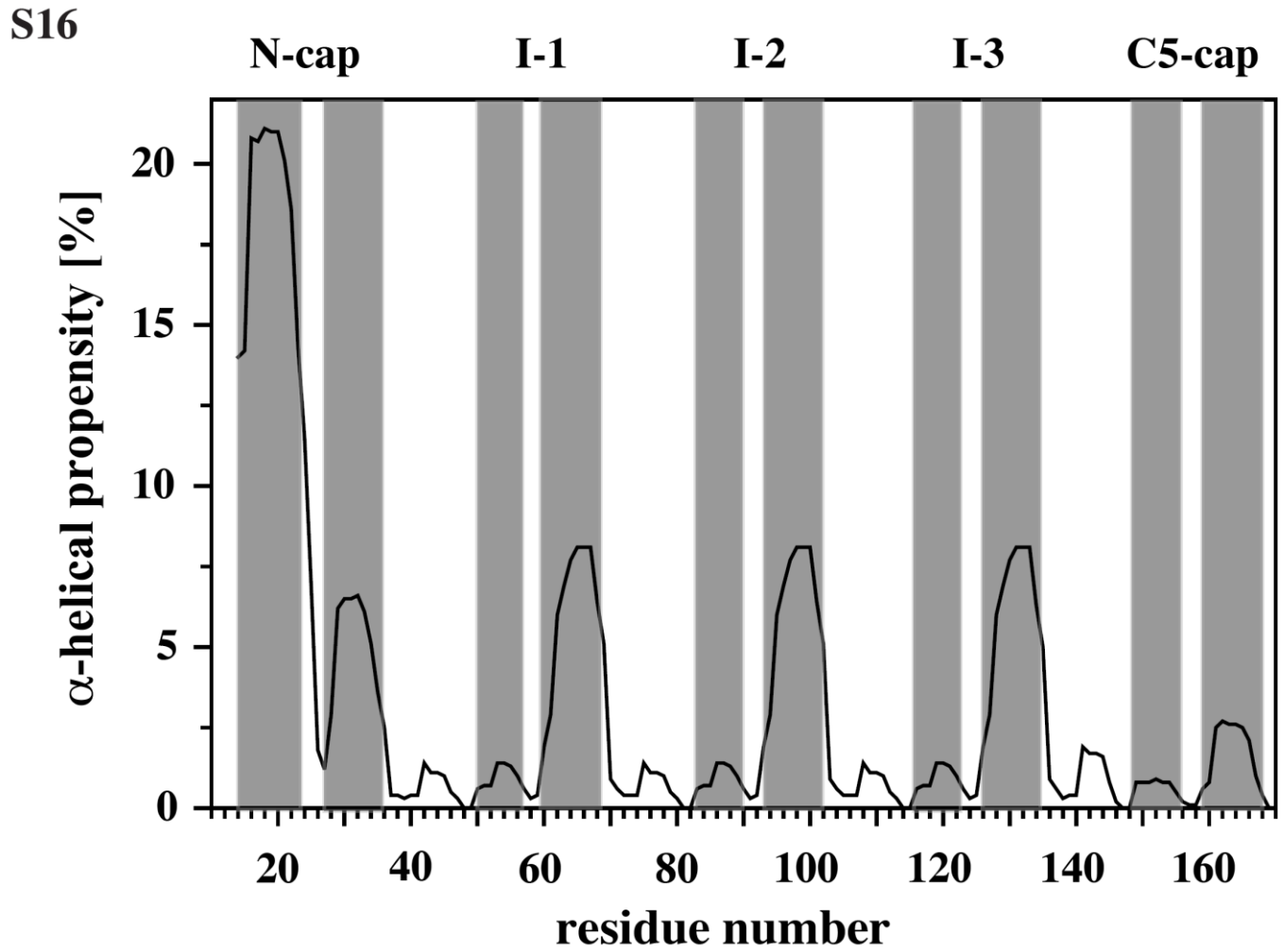


Figure S2.16 Predictions of α-helix for NI_3C_Mut5 content based on the primary sequence using the program AGADIR

3 Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy

Pietro Alfarano[1¥], Gautham Varadamsetty[1¥], Christina Ewald[2], Fabio Parmeggiani[1], Riccardo Pellarin[1], Oliver Zerbe[2], Andreas Plückthun[1*], Amedeo Caflisch[1*]

Published in: Protein Science, September 2012, Volume 21, Isuue 9, pages 1298-314

[1] Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

[2] Department of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

[¥] these two authors contributed equally to the work

*corresponding authors:

Email address of corresponding authors: caflisch@bioc.uzh.ch; plueckthun@bioc.uzh.ch

Keywords: repeat proteins, protein design, structural biology, implicit solvent

3.1 Abstract

A multidisciplinary approach based on molecular dynamics (MD) simulations using homology models, NMR spectroscopy, and a variety of biophysical techniques was used to efficiently improve the thermodynamic stability of armadillo repeat proteins (ArmRPs). ArmRPs can form the basis of modular peptide recognition and the ArmRP version on which synthetic libraries are based must be as stable as possible. The 42-residue internal Arm repeats had been designed previously using a sequence-consensus method. Heteronuclear NMR revealed unfavorable interactions present at neutral but absent at high pH. Two lysines per repeat were involved in repulsive interactions, and stability was increased by mutating both to Gln. Five point mutations in the capping repeats were suggested by the analysis of positional fluctuations and configurational entropy along multiple MD simulations. The most stabilizing single C-cap mutation Q240L was inferred from explicit solvent MD simulations, in which water penetrated the ArmRP. All mutants were characterized by temperature- and denaturant-unfolding studies and the improved mutants were established as monomeric species with cooperative folding and increased stability against heat and denaturant. Importantly, the mutations tested resulted in a cumulative decrease of flexibility of the folded state *in silico* and a cumulative increase of thermodynamic stability *in vitro*. The final construct has a melting temperature of about 85 °C, 14.5 degrees higher than the starting sequence. This work indicates that *in silico* studies in combination with heteronuclear NMR and other biophysical tools may provide a basis for successfully selecting mutations that rapidly improve biophysical properties of the target proteins.

3.2 List of Abbreviations:

2D	two-dimensional
ANS	1-anilino-8-naphthalene sulfonate
ArmRP	Armadillo Repeat Protein
CD	circular dichroism
GdnHCl	guanidine hydrochloride
HSQC	heteronuclear single-quantum coherence
MD	molecular dynamics
NMR	nuclear magnetic resonance
PCR	polymerase chain reaction
RMSD	root mean square deviation
RMSF	root mean square fluctuation
SDS-PAGE	sodium dodecylsulfate polyacrylamide gel electrophoresis
SEC	size-exclusion chromatography

3.3 Introduction

Molecular recognition is a very important aspect of biochemistry and is involved in almost all biological processes. Consequently, it is also the basis of numerous procedures in biological research and biomedical applications. To extend the applications beyond what is possible with antibodies, a number of different protein scaffolds[1-3] were explored over the past two decades for the generation of designed binding molecules using both rational and combinatorial approaches.

While many recognition processes involve the mutual recognition of folded proteins, unstructured regions also play an important role. They frequently occur in linkers and termini of folded proteins, and many posttranslational modifications (e.g. phosphorylation, acetylation, methylation, etc.) are usually within extended regions of proteins. The recognition of unstructured regions of proteins has important applications in proteomics, as proteins frequently get denatured or even need to be unfolded by denaturants or detergents for analysis, such as e.g. for Western blots or protein chips. Additionally, the analysis by mass spectrometry frequently requires a proteolytic digestion, in which the proteins also lose all structural information. The sequence-specific recognition of unfolded proteins or extended regions or termini could thus enable the identification or quantitation of proteins or mutants in a very efficient way, using numerous technologies.

Repeat proteins are made up of several tandem repeats of defined structural units, which create an extended superhelical structure. They are especially attractive for designing binding proteins because of the modular nature of their surface.[2] Several repeat proteins bind peptides, such as HEAT-repeats,[6] Armadillo-repeats[7-10] or TPR-repeats.[11] We found Armadillo repeat proteins (ArmRPs) of particular interest, since they bind a peptide in an extended conformation along a continuous surface contributed to by each module[12], each of which can form contacts to two consecutive amino acids.[13-15]

The ArmRP family received its name when the first member discovered was found to be encoded by the *armadillo* locus, the DNA region that codes for a set of segment polarity genes required during *Drosophila* embryogenesis.[16, 17] This protein is now recognized as the *Drosophila* homolog of β-catenin, involved in Wnt signaling.[18-20] Importin-α is another important member of the family, recruiting the nuclear localization sequence (NLS) in the classical import pathway of cargo molecules into the nucleus.

Armadillo repeats are made up of 42 amino acids formed by three α-helices, named H1, H2 and H3. Helix H3 forms multiple contacts with the bound peptide, amongst them hydrogen bonds from a conserved asparagine residue to main chain peptide bonds. Other side chains on the binding surface provide the specificity for the peptide sequence. Internal repeats have a solvent-accessible surface and two buried surfaces, where they contact neighboring flanking repeats (Figure 3.1). The first and last repeats, called N- and C-terminal capping repeats (or N- and C-caps for short), respectively, have only one buried surface. In case of ArmRPs, the N-terminal cap is shorter than the other repeats, as the N-cap only begins with helix 2.

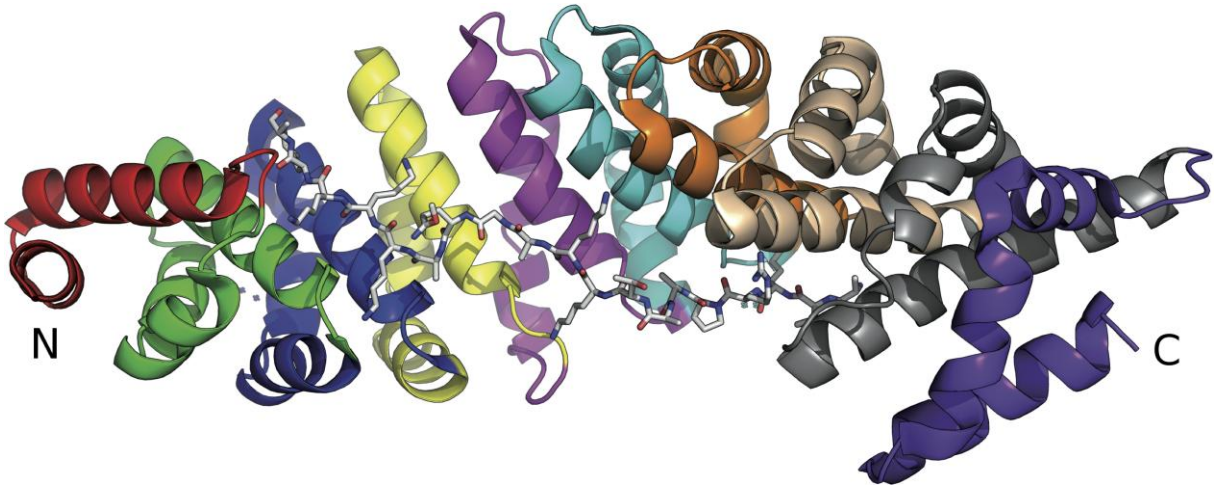


Figure 3.1 An armadillo repeat protein bound to a peptide. Importin-α (PDB accession code: 1EE5)43 in complex with a nucleoplasmin NLS peptide is shown. Every repeat is colored differently and the NLS peptide is in stick representation

Several crystal structures of ArmRPs in complex with different NLSs (a representative one is shown in Figure 3.1) revealed that the NLS peptide runs antiparallel to the direction of the importin-α main chain and that the NLS peptide crosses helix H3 at an angle of approximately 45°. In a first approximation, the complex of the NLS peptide to the ArmRP can be described as an asymmetric antiparallel double helix.

In our efforts to develop ArmRPs with defined binding specificity we initiated a project aimed at creating an ArmRP of utmost stability that will subsequently serve as the scaffold from which libraries are generated to select for specific peptide binding. Previously, Parmeggiani *et al.*[12] designed artificial ArmRPs derived from a consensus sequence, optimizing the hydrophobic core using a computational approach. The consensus sequence had been obtained by multiple sequence alignments of single armadillo repeat modules from both the importin-α and the β-catenin families to generate a unique stable internal module sequence.

The aim of the present study was to further improve the stability of these proteins. Prompted by earlier studies,[12] in which NMR spectra showed markedly better spectra for these proteins at very high pH, we investigated positions of potential electrostatic repulsions at neutral pH in the internal repeats. Moreover, we focused on the optimization of the N- and C-terminal caps. Previous work in designed ankyrin repeat proteins (DARPins) had demonstrated a significant influence of cap engineering on the overall stability of the protein.[21]

While this study was carried out, no crystal structure of a designed ArmRP was available, and thus it was based on homology models largely derived from importin-α. We used implicit solvent molecular dynamics (MD) as well as explicit water MD to assess the fluctuations of different regions in the protein, notably the caps. Based on these simulations, mutants were constructed and experimentally tested. Using a systematic approach optimizing the electrostatics of the internal repeats, and the sequence of the N- and C-terminal caps, proteins could be constructed that are entirely monomeric, posses melting temperatures as high as

85°C, and display biophysical properties as well as NMR spectra characteristic of well-folded and stable proteins at neutral pH.

3.4 Results

The goal of ArmRP engineering is to create a stable scaffold for the generation of libraries as the basis for selecting a new type of sequence-specific peptide binders, where the peptides are bound in an extended conformation. For this purpose it is crucial to create an ArmRP scaffold of utmost stability as a starting point, since we expect that mutations required to achieve binding will inevitably lower the stability of the proteins. We describe here a combination of computational and biophysical approaches to design and characterize the mutants.

Initially, attempts to crystallize the various consensus ArmRP designs had been unsuccessful. Suggestions for modifications of the sequences of the internal repeats came from early NMR studies and homology models. Modifications of the caps were largely derived from MD simulations based on homology models. The MD simulations provided insight into the molecular features that affect structural stability of these proteins. Promising mutants were expressed, and assessed by heteronuclear NMR regarding stability and side chain packing, as well as by thermal and denaturant-induced unfolding observed by optical spectroscopy. The tight interplay of computational techniques with NMR and other biophysical techniques helped to rapidly improve the stability of the ArmRP.

To succinctly describe the proteins with regards to repeat identity and repeat numbers we have introduced a shorthand nomenclature, which should be consulted in Materials and Methods. The protein that was at the start of our studies is termed YM_4A.

3.4.1 Optimization of the internal repeats using heteronuclear NMR spectroscopy

$^{15}N,^{1}H$ heteronuclear NMR spectroscopy represents a suitable tool to investigate the state of folding of small to medium-sized proteins. While other biophysical data have indicated that YM_4A is a well-folded protein[12], spectra recorded at close to neutral pH displayed very broad lines, indicating the presence of conformational exchange processes. Interestingly, when the pH was adjusted to a value of 11, most of the peaks in the 2D-NMR spectrum appeared well-resolved and narrow. Due to accelerated amide exchange at that pH, however, peaks arising from Gly residues, which are mostly located in loops and hence not protected from exchange, disappear **(**Figure 3.2**c)**. As a result signal dispersion in the ^{15}N dimension is limited.

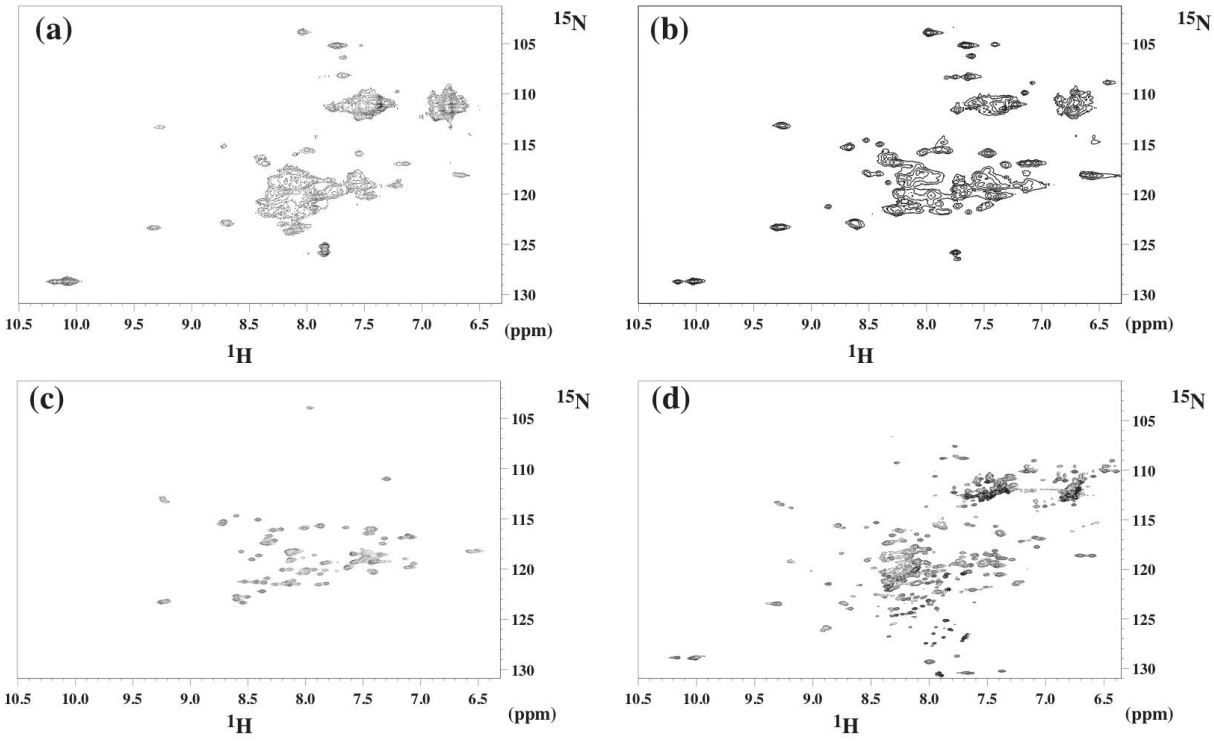


Figure 3.2 (a) to (c): Representative [^{15}N,^{1}H]-HSQC spectra of YM$_4$A recorded at various values of pH: Top left: pH 8.0, top right pH 9.0 and bottom left pH 11.0. Panel (d) displays the spectrum of Y$\bar{\text{M}}_4$A, (= YM$_4$A with the mutations K26Q,K29Q in every repeat = QQ type) at pH 8.0 for comparison

We suspected that the pH dependence of the NMR spectrum may be attributed to the titration of Lys residues, for which side chain pK_as are typically about 10.5. Accordingly, at pH lower than 10 the ε-amino group is charged, resulting in unfavorable side-chain packing. Two Lys residues at position 26 and 29 in each repeat are arranged such that they may form repulsive interactions between the repeats (Figure 3.3). A series of mutants in which, in each repeat, either one of these two Lys residues (data not shown) or both were replaced by Gln **(**Figure 3.2**d)** indicated that the best spectra were obtained when both Lys residues were replaced. The comparison of spectra of YM$_4$A (containing both Lys, thus KK-type) (Figure 3.2**a**) and Y$\bar{\text{M}}_4$A (both mutated to Gln, thus QQ type) (Figure 3.2**d**) at pH 8.0 clearly illustrates the much-improved properties of the QQ-mutant (see Materials and Methods for nomenclature). We would like to emphasize here that no assignments were required at this stage of NMR analysis. In subsequent work the QQ-mutant was used as the scaffold for further optimizations.

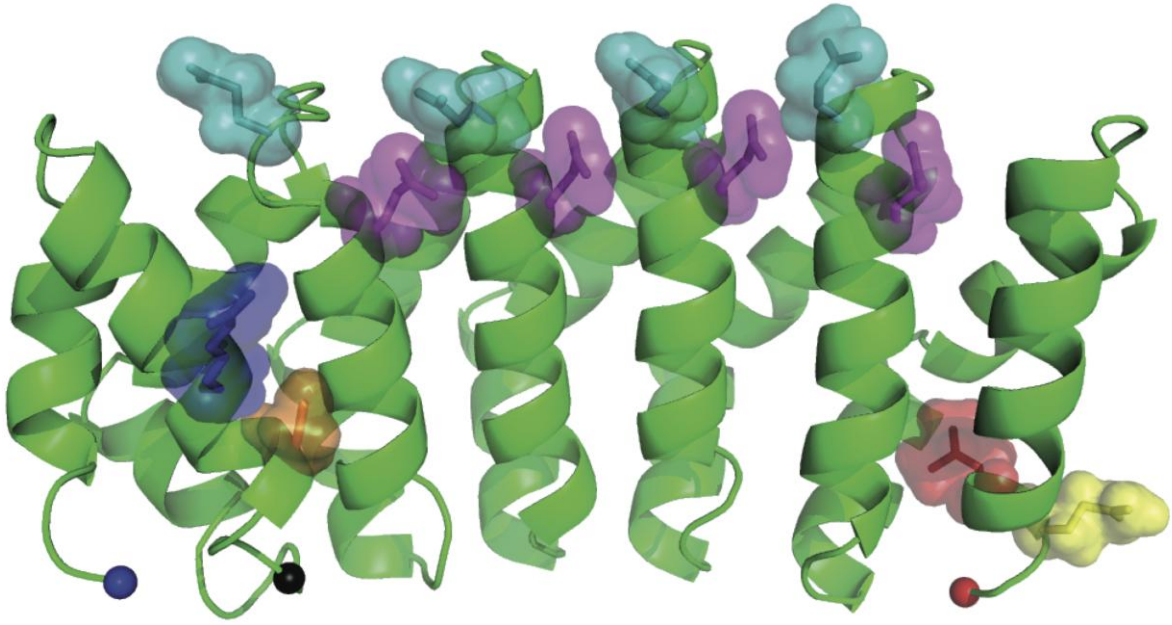


Figure 3.3 $Y_{II}\bar{M}_4A_{II}$ (QQ-type) model displaying the location of the stabilizing mutations as sticks. In the N-cap: R24 (blue) and S27 (orange), the deletion site of R32 is marked by a black ball. In the four internal repeats: Q59, Q62, Q101, Q104, Q143, Q146, Q185 and Q188; the glutamine at position 26 of each repeat is depicted in cyan, the one at position 29 in magenta. In the C-cap: L240 (red) and Q241 (yellow). The locations of the N- and C-terminus of the protein are marked by blue and red balls, respectively

3.4.2 MD simulations suggest mutations at the N-cap and C-cap that result in improved protein stability

A series of MD simulations was carried out to provide suggestions for additional mutations aimed at improving the general stability of the scaffold. An initial explicit water simulation with the model of the original YM_4A (KK-type) provided evidence that the overall fold was preserved during the trajectory. However, two water molecules permeated the interface R4/C-cap close to a buried glutamine (Q240) (Figure 3.4). In the crystal structure of β-catenin (2BCT), this position is occupied by a methionine (M662), which is also buried. Furthermore, the C-cap displayed higher conformational instability than the internal repeats in both the implicit and explicit solvent simulations **(Supp.** Figure S3.2**,** Figure S3.3**, and** Figure S3.4**)**. These simulation results were used to suggest the Q240L mutation at the C-cap, but the Met mutant was also tested experimentally (see below). At the same time, the simulation suggested to mutate the solvent-exposed F241 to glutamine (**Supp.** Figure S3.2). The C-cap containing the mutations Q240L and F241Q is termed "A_{II}".

Since the NMR data indicated that the QQ-mutant displays better side chain packing at neutral pH, a QQ-model ($Y\bar{M}_4A$) was derived from the KK model (YM_4A). The RMSF plot of the KK-model showed high flexibility in both the N- and the C-cap (see Supp. Figure S3.2).

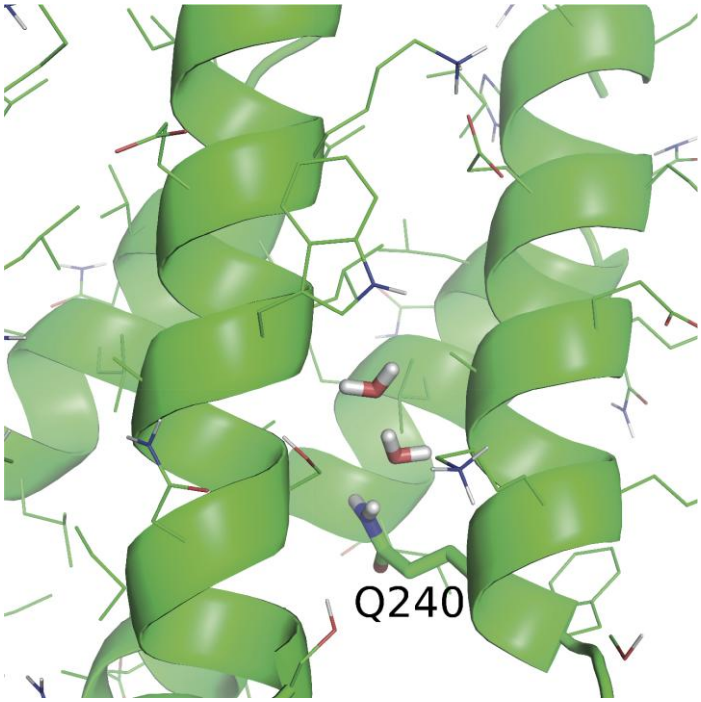


Figure 3.4 Water molecules permeate into the R_4/C interface. In the explicit water simulation of YM_4A, water molecules permeate into the hydrophobic surface between the fourth internal repeat and the C-cap, close to buried Q240

To reduce the flexibility of the N-cap, three mutations were introduced in the QQ-model (Supp. Figure S3.5). Their positions in the sequence are shown in Supp. Figure S3.1. The V24R mutation was introduced to favor an inter-repeat salt bridge with E64, and to remove the solvent exposed V24. The R27 side chain was replaced by Ser, as found in the internal repeats at this position. The loop connecting the N-cap with the first repeat is one residue longer than the ones between the internal repeats (Supp. Figure S3.1). RMSF analysis showed that the backbone R32 is highly flexible. Hence, this residue was deleted to match the length of the loops between internal repeats. The N-cap with all three mutations is termed "Y_{II}" (cf. nomenclature in Materials and Methods).

The mutations investigated in silico are summarized in Table 3.1. To assess the effects of the mutations on the flexibility of the whole protein, the quasiharmonic entropy was calculated (see Material and Methods section). This quantity can be interpreted as an approximation of the configurational entropy. A reduced value corresponds to a reduction of the flexibility and thus to an increase of structural stability. Surprisingly, the average value of the entropy of the $Y\overline{M}_4A$ model is not significantly lower than that of the YM_4A model (Figure 3.5**a**). In contrast, the $Y\overline{M}_4A$-Q240L mutation provides a significant reduction of entropy. Also, the $Y_{II}\overline{M}_4A_{II}$ model, which contains mutations in the N- and C-caps, has the lowest entropy among all variants, in agreement with the NMR spectra (*vide infra*) and biophysical analysis. Furthermore, to investigate the local effect of these mutations, the entropy of the N-cap/R1, R1/R2, R2/R3, R3/R4 and R4/C-cap pairs was calculated (Figure 3.5**b**).

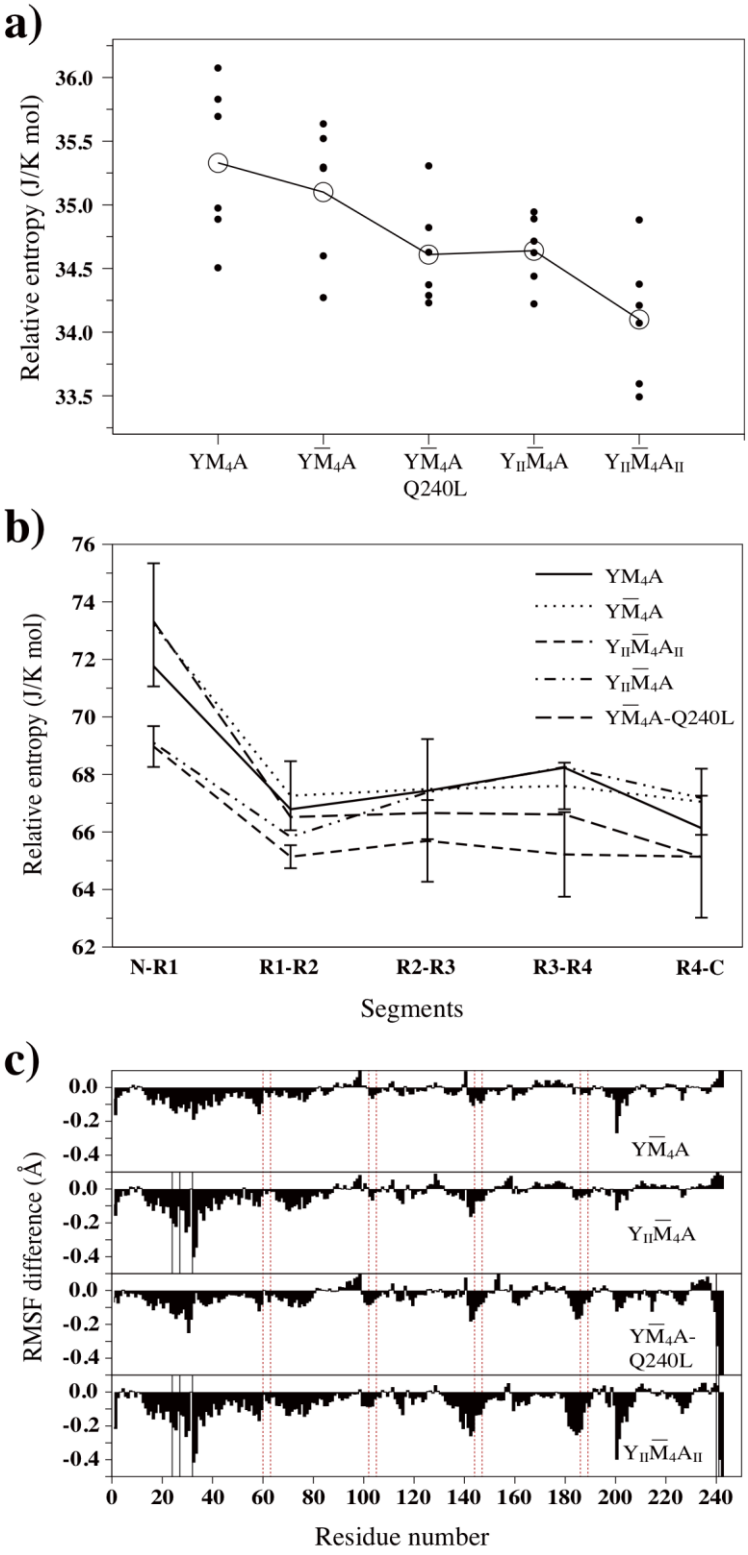


Figure 3.5 Analysis of implicit solvent MD simulations. Panel (a) displays the per-residue quasiharmonic entropy of YM_4A variants. The small filled circles are the results from single MD trajectories and the bigger open circles present their averages. The entropy values are normalized to the number of residues to allow comparing the models with and without the deletion ΔR32 in the N-cap ($Y_{II}\bar{M}_4A$ and $Y_{II}\bar{M}_4A_{II}$). The mutation labels are used as in Table 1. Panel (b) displays changes in quasiharmonic entropy of repeat pairs due to mutations. Error bars represent the standard deviation. Error bars

are only shown for $Y\bar{M}_4A$ and $Y_{II}\bar{M}_4A_{II}$ simulations. The entropy values are normalized to the number of residues in the repeat to allow comparison with the ΔR32 deletion mutants. Panel (c) displays differences in RMSFs of the various $Y\bar{M}_4A$ cap variants, using the RMSF of the YM_4A model as reference. Negative values indicate lower fluctuations relative to the reference. The Lys to Gln mutations introduced in the internal repeats (YM_4A to $Y\bar{M}_4A$ mutations) are indicated by vertical dotted lines, while mutations at the N-cap and C-cap are indicated by vertical solid lines

The trend found when comparing the total entropy of $Y\bar{M}_4A$ and $Y_{II}\bar{M}_4A_{II}$ is reproduced: the quasiharmonic entropy of $Y_{II}\bar{M}_4A_{II}$ is lower than that of $Y\bar{M}_4A$ for all the repeat pairs. Interestingly, when comparing the results for $Y_{II}\bar{M}_4A$ (mutated only in the N-cap) and $Y\bar{M}_4A$-Q240L (mutated only in the C-cap by a single point mutation), the entropy reduction is mainly localized in the mutated capping repeats themselves, without sensibly affecting the internal repeats. An overall decrease of whole and local entropy throughout the whole protein is observed only for $Y_{II}\bar{M}_4A_{II}$, in which both caps are modified.

Similar conclusions can be drawn from a comparison of the root mean square fluctuations (RMSF) (Figure 3.5**c**) between the mutants and the wild-type YM_4A. Mutations at the N- and C-cap measurably reduce the local flexibility of the backbone at the mutation sites. Interestingly, the QQ mutation in the internal repeats, introduced in the $Y\bar{M}_4A$ model, reduces the flexibility of the wild type N-cap. Moreover, $Y_{II}\bar{M}_4A$ and $Y_{II}\bar{M}_4A_{II}$ models have a comparable flexibility, which is lower than that calculated for the YM_4A and $Y\bar{M}_4A$-Q240L models. These observations support the robustness of the method.

To validate the results of the implicit solvent simulations, three independent 80 ns MD simulations with explicit solvent were run for the YM_4A, $Y\bar{M}_4A$, and $Y_{II}\bar{M}_4A_{II}$ models. Therein, the representative of the most populated cluster obtained from the implicit solvent simulations served as starting conformation. The RMSF profiles along the sequence show similar flexibility for implicit and explicit solvent simulations (**Supp.** Figure S3.3 **and** Figure S3.4). Moreover, the three simulations seem to have converged as they individually yield similar RMSF profiles.

Format	Name	Residue 26,29[a]	Mutations
NR_4C	YM_4A	KK	–
NR_4C	$Y\overline{M}_4A$	QQ[b]	K to Q at 60, 63, 102, 105, 144, 147, 186, 189[c]
NR_4C	$Y_{II}\overline{M}_4A$	QQ	QQ + V24R, R27S, ΔR32
NR_4C	$Y\overline{M}_4A$–Q240L	QQ	QQ + Q240L
NR_4C	$Y_{II}\overline{M}_4A_{II}$	QQ	QQ + V24R, R27S, ΔR32, Q240L, F241Q

[a] For the residue numbering of internal repeats see Supporting Information Fig. S1.
[b] These eight mutations are collectively called QQ, and the repeat is termed $\overline{M}$. In combination with a Y_{II} cap these positions are shifted to positions 59, 62, 101, 104, 143, 146, 185, 188 due to the deletion of R32.
[c] Numbering of the entire protein.

Table 3.1 Mutants investigated by MD simulations

To further assess the conformational flexibility, global and local entropies were calculated. Similarly to the implicit water simulations, the global entropy plot **(**Figure 3.6**a)** reveals that $Y_{II}\overline{M}_4A_{II}$ is more rigid than $Y\overline{M}_4A$ and YM_4A. The partial entropy **(**Figure 3.6**b)** shows a trend similar to the one observed in the implicit solvent simulations **(**Figure 3.5**b)**. However, the average conformational flexibility of the YM_4A-model in the N-cap/R1 repeat pair is lower than for $Y\overline{M}_4A$ or $Y_{II}\overline{M}_4A_{II}$. This result is in disagreement with the implicit solvent simulations, where the flexibility of the N-cap/R1 pair of YM_4A is higher than the one of $Y_{II}\overline{M}_4A_{II}$ **(**Figure 3.5**b)**. This discrepancy, as well as the slight increase in the flexibility of the N-cap **(**Figure 3.6 **c),** is in part a consequence of limited sampling in the explicit solvent simulations. For the other repeat dimers flexibility decreases as $YM_4A > Y\overline{M}_4A$, $> Y_{II}\overline{M}_4A_{II}$, in agreement with the implicit solvent simulations.

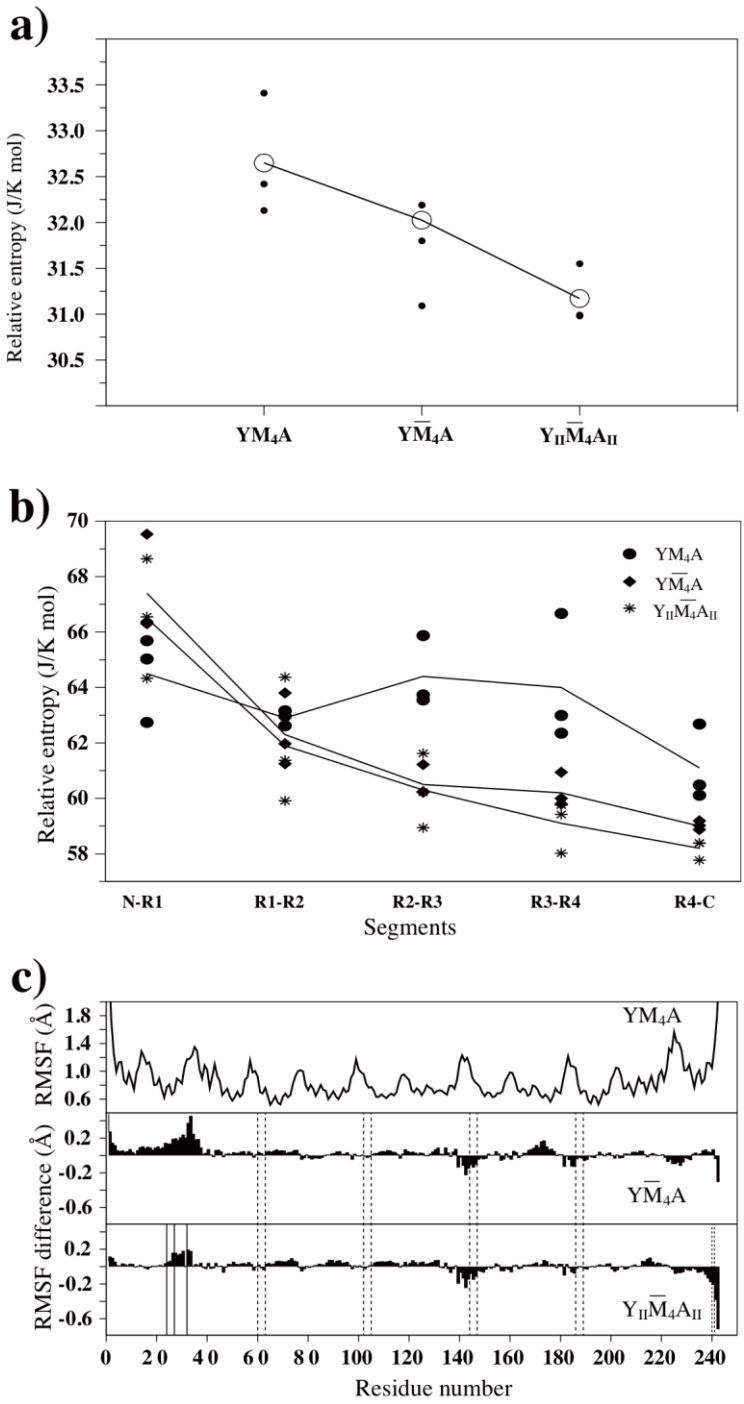


Figure 3.6 Analysis of explicit water MD simulations. Panel (a) Per-residue quasiharmonic entropy derived from explicit water simulations. Three explicit water simulations were run per model. The small filled circles are the individual calculations and the open circles represent the averages. Panel (b) Effect of the mutations on the per-residue quasiharmonic entropy of repeat pairs in explicit water simulations. Panel (c) RMSF comparison of explicit water simulations. The topmost plot is the RMSF plot of the YM_4A-model. Below, the RMSF difference to the YM_4A-model is plotted for every $Y\bar{M}4A$ mutant. Negative values indicate lower fluctuations than for the YM_4A-model. Locations of the Lys to Gln mutations in $\bar{M}$ are indicated by dashed lines, mutations in the N-cap and C-cap by solid lines

It is interesting to analyze the effects of the double mutation R27S and V24R introduced in the Y_{II} cap on the stability of the salt bridges engaged by residue E64. In the original N-cap of YM_4A we observed that E64 strongly interacts with R27. In Y_{II}, as a result of the structural proximity of the newly introduced arginine and the deletion of R27, the salt bridge is formed with R24 **(**Figure 3.7 **right)**.

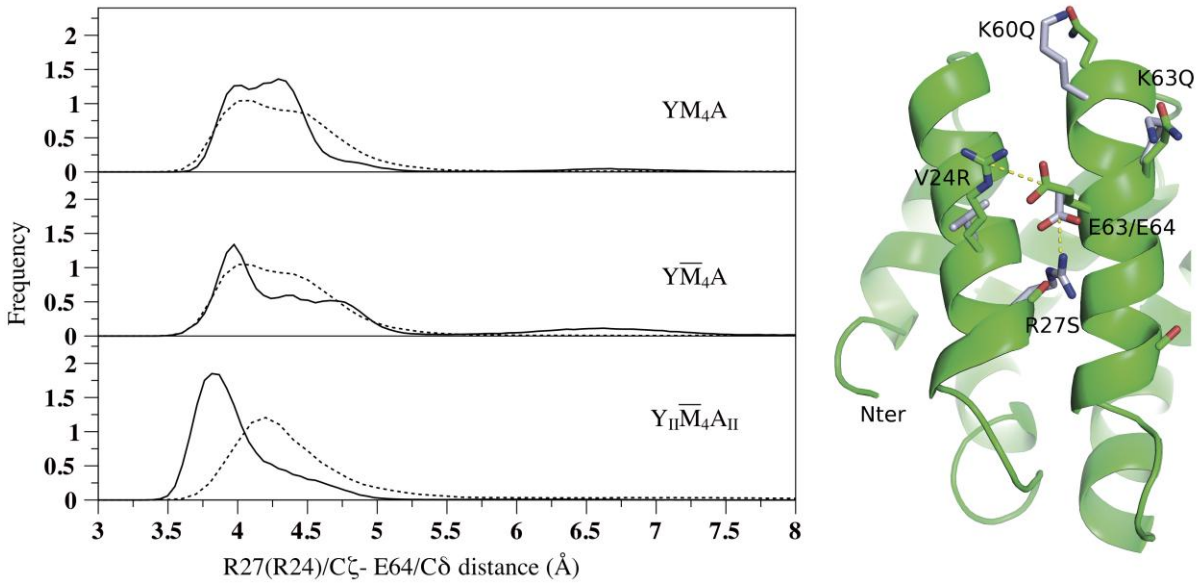


Figure 3.7 Left: Distance distribution of the salt bridge between the N-cap (R27 in YM_4A and $Y\bar{M}_4A$, R24 in $Y_{II}\bar{M}_4A_{II}$) and the first repeat (E64 in YM_4A and $Y\bar{M}_4A$, E63 in $Y_{II}\bar{M}_4A_{II}$). The solid and dotted lines refer to the explicit and implicit solvent simulations, respectively. The salt bridge distance distribution in the case of R27 (top and middle) is less peaked than in the case of R24 (bottom). Right: Ribbon model of the N-cap and the first internal repeat with the salt bridge and mutations. The conformation of $Y_{II}\bar{M}_4A_{II}$ used for starting explicit solvent simulations is depicted in green and side chains from YM4A are colored in white after superposition of the N-cap/R_1 segments. The salt bridges between R27 and E64 (in YM_4A and $Y\bar{M}_4A$) and R24 and E63 (in $Y_{II}\bar{M}_4A_{II}$) are indicated by dashed lines. The side chains of K60 and K63 are shown to illustrate the two residues of the first repeat mutated to glutamines in the $Y\bar{M}_4A$ and $Y_{II}\bar{M}_4A_{II}$ models

We measured the stability of the salt bridges as the ratio of MD snapshots where the distance between the Arg-Cζ and the Glu-Cδ is lower than 4 Å. The comparison between the frequency histograms calculated for the three mutants (Figure 3.7 **left)** reveals that the salt bridge introduced in the $Y_{II}\bar{M}_4A_{II}$ sequence is more stable than the original one. It is worth noting that this result is more pronounced in the explicit than in the implicit solvent simulations. The treatment of the long-range electrostatic interactions and solvation effects are more accurate in the explicit solvent calculations, which may have an influence on the salt-bridge distance range, considering the relatively high solvent exposure of the two side chains involved in the salt bridge.

3.4.3 Biophysical characterization of the M- and M̄-type proteins

Our investigations aimed at constructing very stable consensus ArmRPs have included analysis of the internal repeats, as well as the capping repeats. When changing the Lys residues at position 26 and 29 of the original M repeat (KK-type) to Gln (individually or collectively, but always in all repeats), we found that the QK version led to aggregating molecules and was not pursued further. Both KQ- and QQ-types displayed improved NMR spectra, with the QQ-type (the M̄ repeat) showing the strongest effects (see Figure 3.2**d** above). We thus concentrated on comparing molecules containing M̄-type repeats with those based on the original M-type. This was done in the context of many different cap combinations, which will be discussed below.

To compare the biophysical properties of different ArmRP variants, we carried out expression and solubility tests, CD spectroscopy, thermal and chemical denaturation, and [^{15}N, ^{1}H]-HSQC NMR analysis. All variants were completely soluble and in this respect comparable with the wild-type protein YM_4A. Expression in *E. coli* XL1-blue at 37°C yielded up to 100 mg/l of soluble protein, with similar results for all variants. Immobilized metal-ion affinity chromatography (IMAC) purification yielded pure protein in a single step, as judged by SDS-PAGE (15%). The expected molecular mass values were confirmed by mass spectroscopy.

The CD spectra of all IMAC-purified protein samples **(Fig.** Figure 3.8**a,e,** Figure 3.9**a, Supp.** Figure S3.1**a)** display the expected α-helical secondary structure with minima at 222 nm and 208 nm. The mean residue ellipticity (MRE) of the mutants is similar, but those stabilized in the C-cap show a slightly more pronounced peak at 208 nm **(**Table 3.2**)**.

The CD signal at 222 nm was chosen to monitor thermal and denaturant-induced unfolding. At 10 µM protein concentration, heat denaturation was completely reversible for all proteins (data not shown).

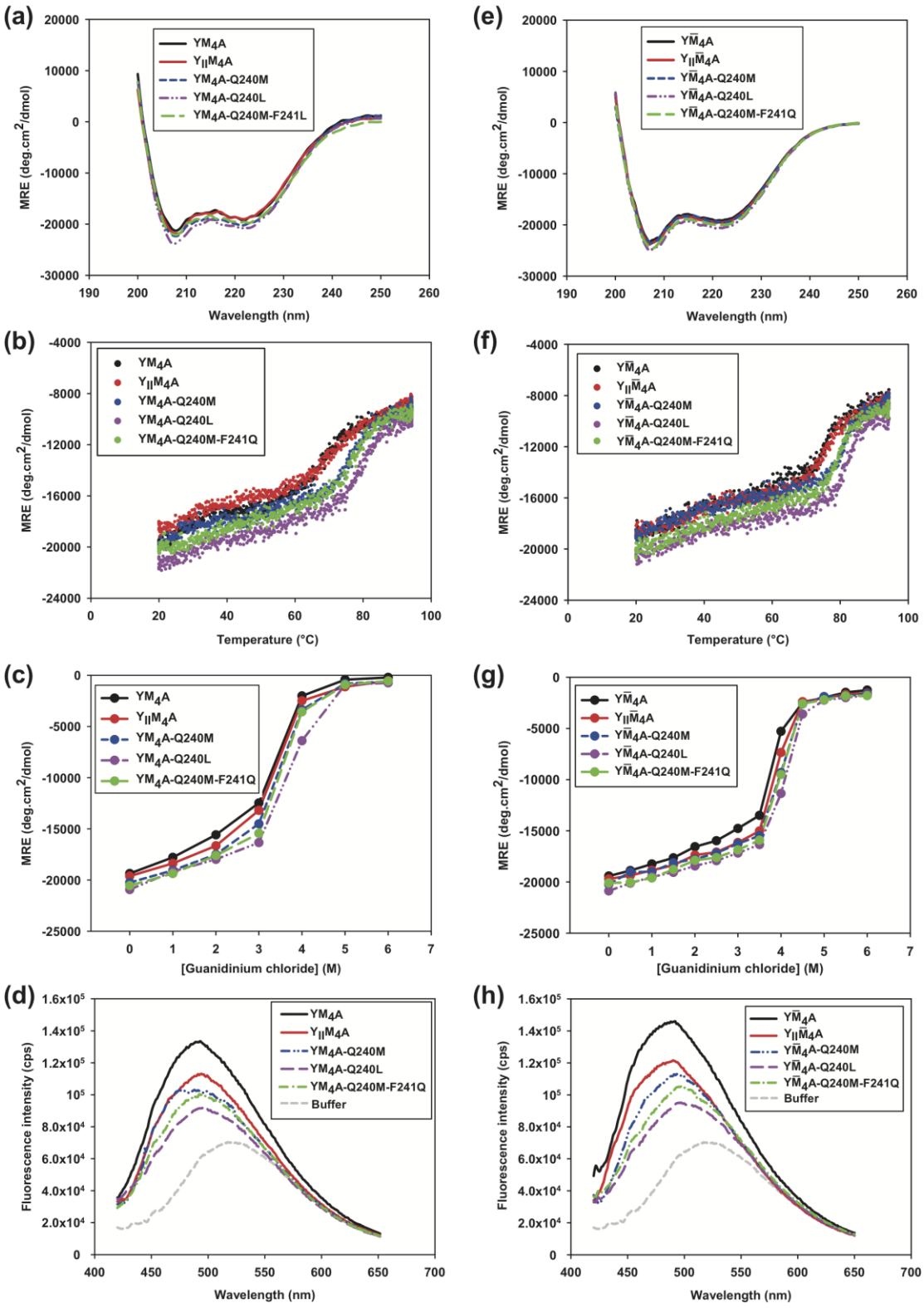


Figure 3.8 Biophysical characterization of designed ArmRP with different consensus repeats M and $\bar{M}$ and cap variants. (a)-(d), YM^{4}A (KK in the internal repeats), (e)-(f), Y$\bar{M}^4$A (QQ in the internal repeats). Identical cap variants have been constructed for both types of internal repeats: Y and Y^{II} for the N-cap; A, A-Q240L, A-Q240M and A-Q240M-F241Q for the C-cap, as indicated in the figure legends. (a),(e) CD spectra; (b),(f) thermal denaturation curves; (c),(g) GdnHCl-induced denaturation curves. The denaturation experiments were followed by CD. The values of MRE at 222 nm are reported. (d),(h) ANS binding. The values without buffer subtractions are shown. The protein concentration was 10 µM

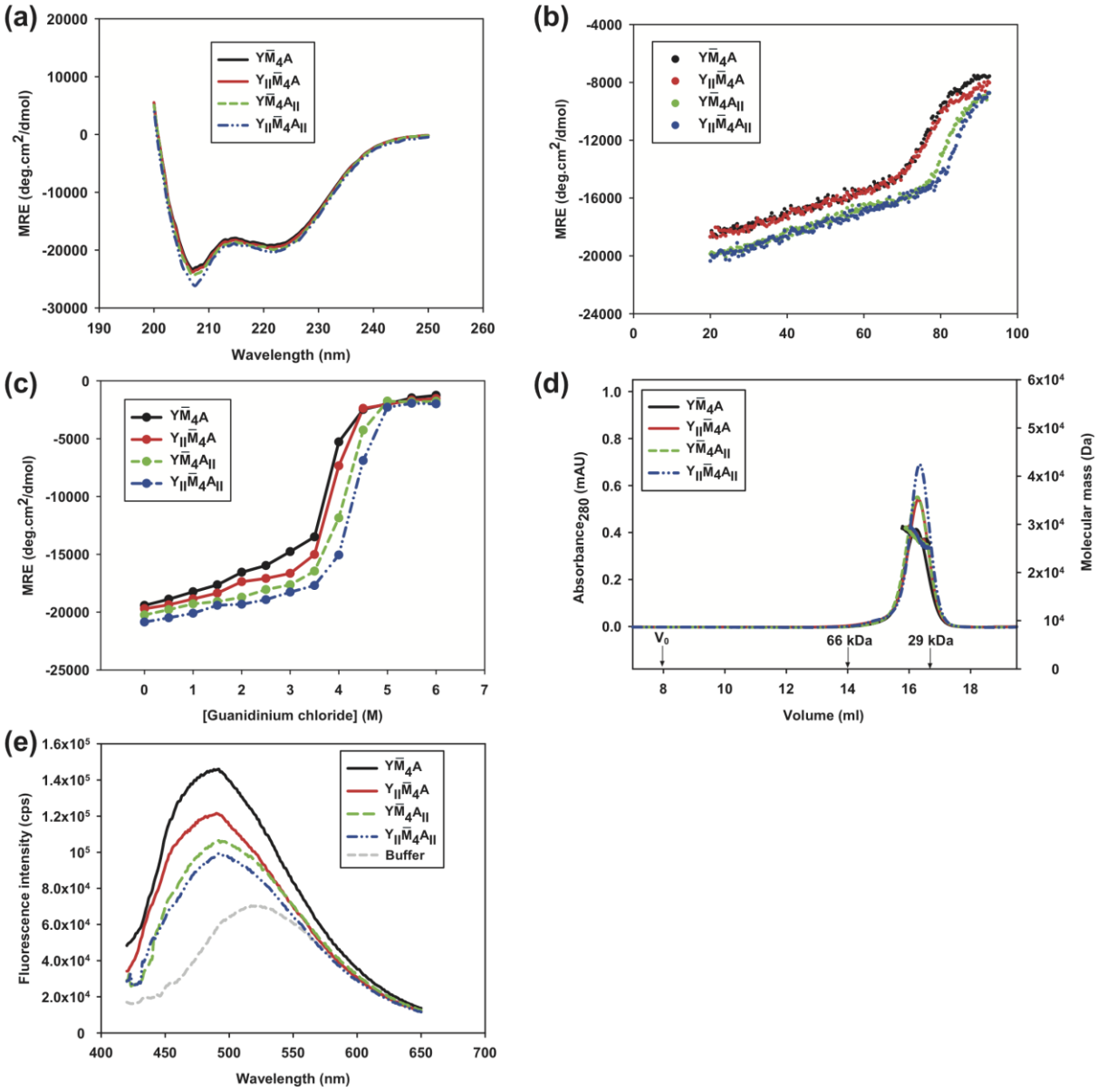


Figure 3.9 Biophysical characterization of designed ArmRPs $YM̄_4A$ and its cap variants ($Y_{II}M̄_4A$, $YM̄_4A_{II}$ and $Y_{II}M̄_4A_{II}$). (a) CD spectra, (b) thermal denaturation curves and (c) GdnCl-induced denaturation curves. The denaturation experiments were followed by CD. The values of MRE at 222 nm are reported. (d) SEC and MALS of designed ArmRPs. The absorbance at 280 nm from SEC is shown on the left y-axis, the calculated MW from MALS on the right y-axis. V0 indicates the void volume of the column. Bovine serum albumin (MW 66 kDa), and carbonic anhydrase (MW 29 kDa) were used as molecular weight markers, and the corresponding elution volumes are indicated by the arrows. (e) ANS binding. The values without buffer subtractions are shown. The protein concentration was 10 µM in a,b,c,e and 30 µM in d

Since both the M and M̄ variants with four internal repeats were modified with analogous capping repeats, we could directly compare the influence of the charge repulsion on a variety

of biophysical parameters (Figure 3.8 Biophysical characterization of designed ArmRP with different consensus repeats M and $\bar{M}$ and cap variants. (**a**)-(**d**), YM4A (KK in the internal repeats), (**e**)-(**f**), $Y\bar{M}_{4A}$ (QQ in the internal repeats). Identical cap variants have been constructed for both types of internal repeats: Y and YII for the N-cap; A, A-Q240L, A-Q240M and A-Q240M-F241Q for the C-cap, as indicated in the figure legends. **(a**),(**e**) CD spectra; (**b**),(**f**) thermal denaturation curves; (**c**),(**g**) GdnHCl-induced denaturation curves. The denaturation experiments were followed by CD. The values of MRE at 222 nm are reported. (**d**),(**h**) ANS binding. The values without buffer subtractions are shown. The protein concentration was 10 ∝M, Table 3.2). For all investigated $Y\bar{M}_4A$ constructs, regardless of the caps, the melting temperature is 4-5°C higher than for the corresponding YM_4A constructs. Similarly, the midpoint of GdnHCl denaturation is 0.2-0.4 M higher. This indicates that the removal of the charge repulsion within the internal ArmR is clearly stabilizing the protein. These results also mean that the effect of the cap variants is quite independent of the internal repeats, thus offering two independent and additive measures to increase stability in ArmR proteins.

We also compared the hydrophobicity of the proteins by evaluating the binding to the fluorescent dye 1-anilino-8-naphthalene sulfonate (ANS) that binds to solvent-exposed hydrophobic patches or to pockets of molten-globule state proteins.[22] When comparing YM_4A and $Y\bar{M}_4A$ constructs with the same caps, ANS binding was very similar (Figure 3.8**d and h**). The caps themselves, however, do influence ANS binding (see below).

3.4.4 Biophysical characterization of various cap mutants allows identifying mutants with much improved stability

The MD simulations have suggested a set of mutations in the caps that should increase the stability of the protein. For validation the new cap variants were constructed in designed ArmRP with both types of internal repeats, YM_4A and $Y\bar{M}_4A$. The proteins were expressed, purified and characterized biophysically. All three N-cap mutations were introduced at once to create the second generation "Y_{II}" N-cap: V24R, R27S and ΔR32 (a deletion mutant). Mutations in the C-cap were investigated individually (Q240L or Q240M), and as double mutant (Q240L/F241Q, denoted as A_{II}) **(**Table 3.1Table 3.2**)**. All mutations were tested in the context of the M and the $\bar{M}$ series, and some mutations were also tested in the $Y\bar{M}_3A$ format **(**Table 3.2**)**.

The thermal stabilities of the cap variants were compared to the respective precursor proteins YM_4A (see Figure 3.8**b**) and $Y\bar{M}_4A$ (Figure 3.8**f**). All proteins displayed a significant slope prior to the main transition and an indication for a cooperative denaturation step at higher temperature.

Constructs[a]	Type	Residues (repeats)[b]	pI[c]	MW_{calc} (kDa)[d]	Oligom. State[e]	MW_{obs} (kDa)[f]
YM_4A	M	253 (6)	4.5	27.1	Monomer	32.3
$Y_{II}M_4A$	M	252 (6)	4.5	27.0	n.d.	n.d.
YM_4A-Q240M	M	253 (6)	4.5	27.1	n.d.	n.d.
YM_4A-Q240L	M	253 (6)	4.5	27.1	n.d.	n.d.
YM_4A-Q240M-F241Q	M	253 (6)	5.1	27.1	n.d.	n.d.
$Y\overline{M}_4A$-Q240M	$\overline{M}$	253 (6)	4.5	27.1	Monomer	32.5
$Y\overline{M}_4A$-Q240L	$\overline{M}$	253 (6)	4.5	27.1	Monomer	32.5
$Y\overline{M}_4A$-Q240M-F241Q	$\overline{M}$	253 (6)	5.1	27.1	Monomer	32.2
Cap combinations						
$Y\overline{M}_4A$	$\overline{M}$	253 (6)	4.5	27.1	Monomer	32.3
$Y_{II}\overline{M}_4A$	$\overline{M}$	252 (6)	4.5	27.0	Monomer	31.6
$Y\overline{M}_4A_{II}$	$\overline{M}$	253 (6)	4.5	27.1	Monomer	31.4
$Y_{II}\overline{M}_4A_{II}$	$\overline{M}$	252 (6)	4.5	26.9	Monomer	31.2
$Y\overline{M}_3A$	$\overline{M}$	211 (5)	4.7	22.8	Monomer	27.5
$Y_{II}\overline{M}_3A$	$\overline{M}$	210 (5)	4.6	22.6	Monomer	28.0
$Y\overline{M}_3A_{II}$	$\overline{M}$	211 (5)	5.2	22.7	Monomer	27.1
$Y_{II}\overline{M}_3A_{II}$	$\overline{M}$	210 (5)	4.6	22.6	Monomer	27.5

[a] Constructs in boldface have been studied by MD simulations (see Table I).
[b] The number of residues includes the $MRGSH_6$ tag; the number of repeats i
[c] Isoelectric point (pI).
[d] Molecular weight calculated from the sequence; masses were confirmed by
[e] Oligomeric state as indicated by multiangle static light scattering.
[f] Observed molecular weight as determined by SEC.
[g] Ratio between observed and calculated molecular weight $MW_{obs/calc}$.
[h] Mean residue ellipticity at 222 nm expressed as $deg \cdot cm^2/dmol$.
[i] T_m observed in thermal denaturation measured by CD.
[j] Difference in T_m relative to YM_4A.
[k] Midpoint of transition in GdnHCl-induced denaturation measured by CD.

Table 3.2 Biophysical properties of designed ArmRPs with different capping repeats.

The modified N-cap in $Y_{II}\overline{M}_4A$ results in a T_m of 77.5°C that is 1.5°C above the transition midpoint of $Y\overline{M}_4A$ wild-type (i.e. T_m = 76°C), suggesting that the N-cap engineering was successful, although its contribution to overall stability is only modest (Table 3.2, Figure 3.8**b, f**).

For the C-cap, the replacement of Gln-240 by a hydrophobic residue resulted in a significant increase in stability to 80°C or 82.5°C for $Y\overline{M}_4A$-Q240M or $Y\overline{M}_4A$-Q240L, respectively, compared to $Y\overline{M}_4A$ **(**Table 3.2**)**. Stability can be further improved by additionally mutating Phe-241 to Gln, with $Y\overline{M}_4A$-Q240M-F241Q and $Y\overline{M}_4A$-Q240L-F241Q (also called $Y\overline{M}_4A_{II}$) displaying transition temperatures of 81°C or 83°C, respectively.

We also investigated unfolding induced by GdnHCl (Figure 3.8**c, g**). All proteins displayed cooperative denaturation in these equilibrium-unfolding experiments. The transition

point for the curves shifted to higher GdnHCl concentrations for the C-cap mutants Q240M, Q240L, and Q240M-F241Q, both in the YM_4A and the $Y\overline{M}_4A$ format. On the other hand, the transition of constructs with the original Y N-cap was almost identical with those carrying the Y_{II} N-cap, again both in the YM_4A and the $Y\overline{M}_4A$ format **(**Figure 3.8**c, g**). The most significant shift in the transition midpoint was observed for the Q240L mutation in the C-cap, and this could again be improved further by additionally mutating Phe-241, to result in Q240L-F241Q (also called YM_4A_{II} or $Y\overline{M}_4A_{II}$).

Similar to the results for heat denaturation, equilibrium denaturation by GdnHCl revealed that the influence of the N-cap engineering is rather minor (cf. YM_4A with $Y_{II}M_4A$ or $Y\overline{M}_4A$ with $Y_{II}\overline{M}_4A$), while the effect of the C-cap mutation is very significant, with the single mutation Q240L increasing the midpoint of $Y\overline{M}_4A$ from 3.7 M to 4.2 M GdnHCl **(**Table 3.2**)**, and the double mutation present in $Y\overline{M}_4A_{II}$ even to 4.25 M GdnHCl.

The purified proteins differ slightly in their running behavior when analyzed by SDS-PAGE **(Supp.** Figure S3.6**)**. Remarkably, the C-cap mutation Q240L and the double mutations Q240L-F241Q present in the A_{II} cap are characterized by a higher mobility in SDS-PAGE, both the in the context of the original M-type and of the $\overline{M}$-type, while N-cap mutations have a smaller effect. This faster running behavior suggests a higher compactness of these proteins and/or an incomplete unfolding by SDS.

The consensus-designed YM_4A and $Y\overline{M}_4A$ and their cap variants display different behavior in ANS binding experiments. The difference between the curves of corresponding constructs differing only by the Q240L mutation **(**Figure 3.8**d and h)** indicates that this mutation in the C-cap reduces the hydrophobic solvent-exposed surface or accessible interface. The mutation probably stabilizes the hydrophobic core indicated by the increase in the midpoint of transition both in thermal and GdnHCl- induced denaturation (Figure 3.8**b-c, f-g**).

3.4.5 Biophysical characterization of cap combinations

Having established that the Y_{II} N-cap and the A_{II} C-cap variants result in the highest improvements in stability, it became of interest to test whether the observed effects are additive or even synergistic. We thus generated the combinations $Y_{II}M_4A_{II}$ and $Y_{II}\overline{M}_4A_{II}$ and investigated their properties in more detail.

The stability of the combined cap mutant $Y_{II}\overline{M}_4A_{II}$ was assessed by thermal and GdnHCl-induced denaturation (Figure 3.9**b,c**). $Y_{II}\overline{M}_4A_{II}$ possesses a melting temperature of T_m= 85.5°C (Figure 3.9**b,** Table 3.2). Compared to the variant with the original N-cap ($Y\overline{M}_4A_{II}$), the increase in stability is 2.5°C, or 8°C compared to the variant with the original C-cap ($Y_{II}\overline{M}_4A$). This demonstrates that most of the additional stability is contributed by the engineered C-cap. These data also reveal that the cap improvement is additive to a first approximation, suggesting negligible cooperative interactions throughout the whole protein. In summary, when the engineered Y_{II}- and A_{II}- caps are combined, an increase in the melting point by almost 10°C is observed, compared to $Y\overline{M}_4A$, and almost 15°C are obtained relative to the original YM_4A (T_m = 71°C), demonstrating the success of our engineering efforts (Table 3.2).

In the GdnHCl-induced unfolding experiments of $Y\bar{M}_4A_{II}$ and $Y_{II}\bar{M}_4A_{II}$ the transition point for the curves are shifted to higher GdnHCl concentrations, compared to $Y\bar{M}_4A$, while the transition of $Y_{II}\bar{M}_4A$ was almost superimposable with that of $Y\bar{M}_4A$ **(**Figure 3.9**c**). The highest shift in the transition point was observed for $Y_{II}\bar{M}_4A_{II}$, consistent with the data obtained in temperature-induced unfolding (Table 3.2). Again, the effect was only modest for N-cap engineering ($Y\bar{M}_4A \rightarrow Y_{II}\bar{M}_4A$ and $Y\bar{M}_4A_{II} \rightarrow Y_{II}\bar{M}_4A_{II}$ shifted by 0.1 or 0.15 M GdnHCl, respectively), and more pronounced for C-cap engineering ($Y\bar{M}_4A \rightarrow Y\bar{M}_4A_{II}$ and $Y_{II}\bar{M}_4A \rightarrow Y_{II}\bar{M}_4A_{II}$ shifted by 0.55 or 0.6 M GdnHCl, respectively), and the effects were again additive to a first approximation.

The difference between the curves of $Y\bar{M}_4A$ and $Y_{II}\bar{M}_4A_{II}$ in the ANS binding experiments demonstrates that the cap mutations reduce the solvent-exposed hydrophobic surface **(**Figure 3.9**e)**. SEC-MALS analysis displayed single symmetric peaks for all variants, and the determined mass indicates a monomeric state (Figure 3.9**d**). The smaller elution volume than for the globular proteins of the standard (Table 3.2**)** is thus almost certainly due to the elongated shape of the molecules. Similar trends and results **(Supp.** Figure S3.7**)** were observed when the cap mutations were introduced into $Y\bar{M}_3A$, and are summarized in Table 3.2.

Considering the inherent error in the stability measurements, the data are consistent with a fairly constant gain in stability while going from $Y\bar{M}_3A$ to $Y\bar{M}_4A$, independent of the caps. In summary, we could increase the stability of designed ArmRPs by four *additive* components: by engineering the N-cap, the C-cap, and electrostatics of the internal modules ($M \rightarrow \bar{M}$), and by increasing the number of internal repeats.

3.4.5.1 Heteronuclear NMR allows to rank $Y\bar{M}_3A$ and $Y\bar{M}_4A$ cap mutants according to their conformational stability

The potential of (heteronuclear) NMR to judge the conformational stability of proteins has been increasingly exploited in the course of structural genomics projects.[23] In this study, 1D ^{1}H-NMR spectra of all proteins were recorded (data not shown) in order to preliminarily evaluate the influence of different mutations or combinations of mutations in the capping repeats of $Y\bar{M}_3A$ and $Y\bar{M}_4A$ with respect to conformational rigidity. Wild-type consensus proteins and their mutants were ranked according to signal dispersion in the amide- and methyl-region as well as the linewidth of their proton resonances. A subset of these, namely the original consensus proteins $Y\bar{M}_3A$ and $Y\bar{M}_4A$, and the improved cap mutants Y_{II} and A_{II} described above **(**Table 3.1**)**, which all appeared to be well structured in 1D proton NMR spectra, were expressed in uniformly ^{15}N-labeled form and analyzed using [^{15}N,^{1}H]-HSQC spectra. Since preliminary work (data not shown) had revealed that the single Gln mutants (QK and KQ for pos. 26 and 29 in the M-repeats) displayed less favorable properties, they were not further pursued here.

The repetitive nature of the sequence and the inherently reduced signal dispersion in purely α-helical proteins is expected to result in limited signal dispersion **(**Figure 3.10**)**. This feature is seen particularly well in the center of the spectrum (see the region between 7.9 and 8.4 ppm in the ^{1}H dimension in Figure 3.10). Due to overlap of peaks fewer than the expected

number of peaks were usually observed, e.g. for $Y_{II}\bar{M}_3A_{II}$ 170 out of the expected 192 cross-peaks were visible. Nevertheless, signal dispersion is remarkably good, and significantly further improved in the cap mutants, as compared to the original $Y\bar{M}_3A$ and $Y\bar{M}_4A$. The line widths suggest that all proteins are monomeric, in agreement with results obtained by size-exclusion chromatography and MALS experiments (Figure 3.9**d** and **Supp.** Figure S3.7**d)**. Interestingly, the effects due to the C-cap mutations Q240L and F241Q (A_{II}) again are stronger than those of the N-cap mutations (V24R, R27S and the deletion of R32; Y_{II}), a feature that was also observed in the MD simulations and in the biophysical characterization of the mutants. The combination of N- and C-cap mutations displays a synergistic effect, resulting in the best signal dispersion and comparably narrow lines for $Y_{II}\bar{M}_3A_{II}$ (Figure 3.10).

Spectra for the $\bar{M}_4$ series displayed similar trends although the increase in line width due to the larger size was significant (**Supp.** Figure S3.8). Again, the results for the $Y_{II}\bar{M}_4A_{II}$ construct are consistent with the observations from equilibrium unfolding studies and the MD simulations.

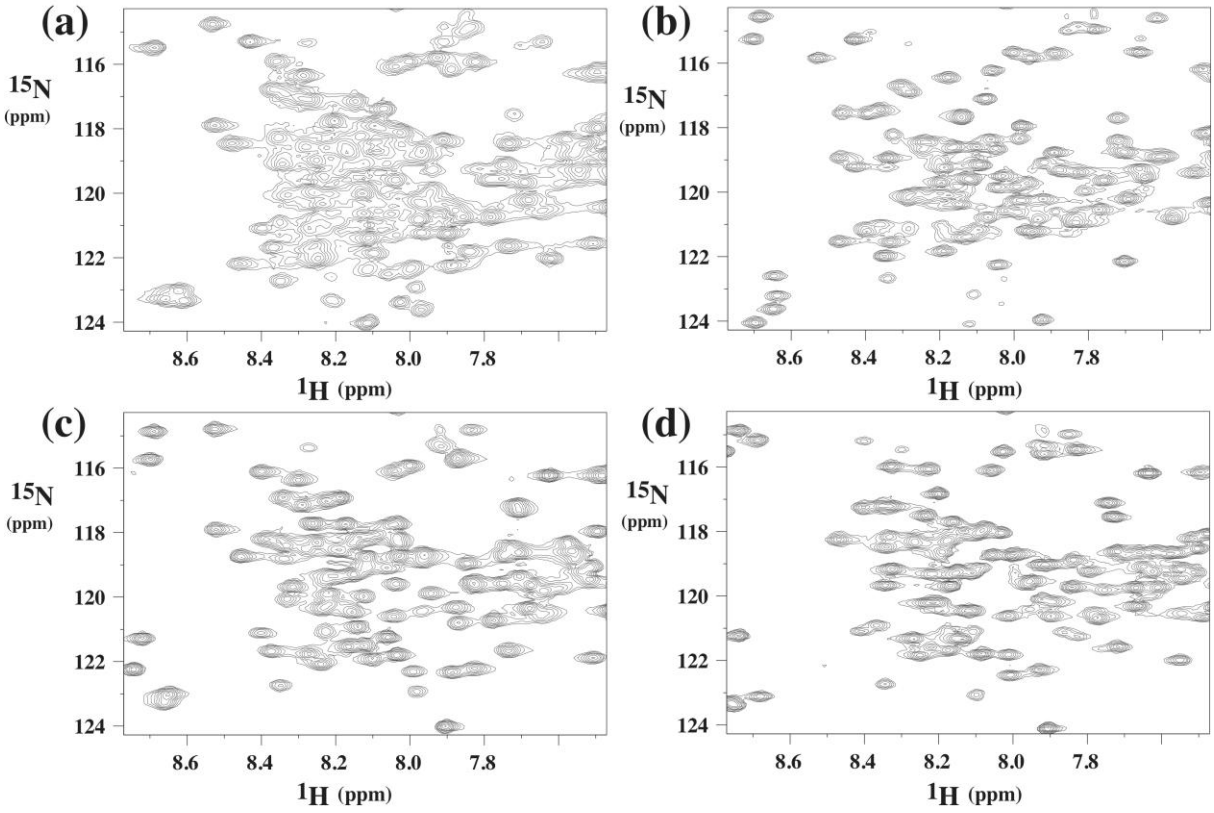


Figure 3.10 [^{15}N, ^{1}H]-HSQC spectra of designed ArmRP $Y\bar{M}_3A$ (a) **and its cap variants. $Y_{II}\bar{M}_3A$** (b), **$Y\bar{M}_3A_{II}$** (c) **and $Y_{II}\bar{M}_3A_{II}$** (d) **at pH 7.4. All spectra were recorded at 310 K in 50 mM phosphate buffer, and 150 mM NaCl. The protein concentration was 0.5 mM**

3.5 Discussion

Engineering of proteins for increased stability is a prerequisite for using them as a starting point for randomization, as is needed in the creation of libraries for the selection of binding molecules. While we have used consensus engineering initially[12] and have already applied a computationally guided optimization of the hydrophobic core of the internal ArmRs, the stability of the resulting proteins was still unsatisfactory.

Herein we have developed a method in which stability of proteins is improved using a rational approach that results in the expression of only a few mutants but nevertheless very effectively increased the stability. The approach uses MD simulations based on homology models of the repeat proteins to provide important information for suggesting the mutations. Furthermore, heteronuclear NMR helped to detect a charge repulsion problem in the internal repeats that resulted in destabilization of the protein and improper side-chain packing. In general, NMR was useful to correctly rank the stability of proteins even in the absence of any backbone assignments.

Improvements were obtained by removal of electrostatic repulsions within the internal repeats. However, cap re-engineering guided by MD simulations made the largest contribution. NMR measurements and a variety of biophysical measurements confirmed that the newly designed N-and C-cap mutants are significantly more stable and better structured. The largest increase in stability is due to modifications of the C-cap, and in particular to the Q240L mutation, as demonstrated by thermal and chemical denaturation experiments **(**Figure 3.8**b,c,f,g)**. Furthermore, NMR measurements confirm that the newly designed Y_{II} and A_{II} mutants (as present in $Y_{II}\overline{M}_4A_{II}$ and $Y_{II}\overline{M}_3A_{II}$) are significantly more stable and better structured than the corresponding initial constructs $Y\overline{M}_3A$ and $Y\overline{M}_4A$. The reduced line width observed in the [^{15}N,^{1}H]-HSQC spectra is most likely due to better packing of side chains. Hence, the NMR data are in good agreement with predictions from MD simulations and results from thermal and chemical unfolding experiments and the ANS-binding behavior of the tested proteins. The more stable A_{II} cap therefore “couples” better to the rest of the protein. In summary, the weak link in the artificially designed original C-cap has been strengthened by our engineering, inspired by MD simulations.

Apparently, the better packing of the C-cap against internal repeats due to this modification prevents local unfolding events that may eventually trigger complete unfolding. This observation is supported by results from our previous study of proteins with Ankyrin repeats, in which we observed a similar influence of the stability of capping repeats on the overall protein stability.[21, 24, 25] In fact, the Ising model predicts that the stability of these proteins arises from mutual stabilization of neighboring repeats,[24-27] and these effects are therefore expected to propagate throughout the entire protein. In principle, the stabilization should be similar for all repeats, but our experience has shown that the potential for optimization is the largest for the capping repeats. A very important result of this study is that both caps could be re-engineered *independently*, and that improvements resulting from the modified caps were additive (and perhaps synergistic) to a first approximation. In MD simulations a lower flexibility of the internal repeats was seen only when both caps are mutated, otherwise stabilization remained a local effect.

This work highlights several strategies for improving the stability of repeat proteins. Similar to the original work by Parmeggiani,[12] where the hydrophobic core of the internal repeats has been optimized, here the hydrophobic core of the caps was improved. Additionally, electrostatic repulsions in the internal repeats were found to be a main contributor, as shown in the conversion of the M-type modules to the $\overline{M}$-type modules. In the caps an additional attractive interaction has conferred more rigidity. Because of the modular nature of repeat proteins, the improvements in the internal repeats and the caps can easily be combined. Finally, the very simple addition of more internal repeat increased the stability of the repeat proteins.

In summary, this work has brought consensus-designed armadillo repeat proteins through various generations of engineering to a point that they can now form the basis of libraries for the construction of sequence-specific peptide binders. Evolved ArmRP, based on the $Y\overline{M}_4A_{II}$ sequence and engineered for binding to neurotensin, allowed their successful study by NMR due to the much-improved stability of these mutations. In contrast, initial work on mutants based on the YM_4A design was unsuccessful because the derived proteins rapidly oligomerized and/or precipitated (data not shown). Our experience therefore underlines the value of optimizing the basic skeleton before introducing mutations for ligand binding. The present work indicates that this optimization process can be guided and accelerated by computational studies.

3.6 Materials and Methods

3.6.1 Nomenclature

The consensus armadillo proteins investigated here consist of an N-terminal capping repeat, derived from *y*east importin-α, termed "Y". It is followed by several consensus repeats, which are termed "M" and have been described previously,[12] and their number in the protein is given as a subscript. Finally the protein contains a C-terminal capping repeat, which was *a*rtificially designed,[12] termed "A". A protein with four internal repeats is thus called YM_4A. To indicate improved (e.g. second generation) versions of the capping repeats, the caps are labeled with a roman numeral, e.g. $Y_{II}M_4A_{II}$.

The sequences of Y-type N-cap, M-type internal repeat and A-type C-cap are shown in **Supp.** Figure S3.1. In the present study, we have also investigated the effect of mutating Lys26 and Lys29 to Gln (numbering of individual repeats), individually or in combination. This was always done for every repeat in a protein at once. We thus refer to these two residues in the single-letter code: the original M-type internal repeat[12] with Lys26 and Lys29 is thus referred to as the KK-type, and from this QK, KQ and QQ have been generated. Thus, the YM_4A (QQ-type) sequence carries mutations of lysine residues 60, 63, 102, 105, 144, 147, 186 and 189 (numbering based on the whole protein). To abbreviate the nomenclature further, we refer to the original M-repeat (KK-type) as "M" and the newly engineered M-repeat (QQ-type) as "$\overline{M}$". Thus, the proteins would be termed, e.g. YM_4A and $Y\overline{M}_4A$.

When it is necessary to specify an individual internal repeat, R*i* stands for the i^{th} internal repeat, and R*i*-R*j* stands for the repeat pair composed by the i^{th} to j^{th} repeats.

3.6.2 MD simulations

Langevin dynamics simulations were performed at 300 K using the program CHARMM[28] and the implicit solvent FACTS.[29] The protein was modeled according to the united atom CHARMM PARAM19 force field.[30] The protonation state of the side chains was chosen to reproduce pH 7.4 of the CD and NMR experiments: aspartate and glutamate side chains as well as the C-terminal carboxyl group were negatively charged, lysine and arginine side chains together with the N-terminal amino group were positively charged and histidine residues were kept neutral. All bonds between hydrogen and heavy atoms were constrained using SHAKE,[31] allowing an integration step of 2 fs. Different initial random velocities were assigned to every simulation. Unless differently specified, each simulation consisted of three phases: 0.2 ns heating, followed by 0.4 ns equilibration and 30 ns production. About 10.5 hours on a core of a XEON 5410 Quadcore CPU running at 2.33 GHz are required for a 1 ns trajectory of the KK model (nearly 2220 atoms).

Explicit solvent MD simulations were performed at 300 K using the program CHARMM. The protein was modeled according to the all-hydrogen CHARMM force field (PARAM22 with CMAP correction)[32, 33] and TIP3P water model[34] with the same protonation state discussed above. The protein was inserted into a water-filled orthorhombic box whose dimensions were determined such that each atom of the protein had at least 13 Å distance from the boundary. Chloride and sodium ions were added to neutralize the total charge of the system at a concentration of 200 mM. To avoid finite-size effects, periodic boundary conditions were applied. Different initial random velocities were assigned to every simulation. Coulombic and van der Waals interactions were calculated up to a cutoff distance of 12 Å, while long-range electrostatic effects were accounted for by the Particle Mesh Ewald summation method.[35] The temperature was kept constant by the Nosé-Hoover thermostat[36, 37] while the pressure was held constant at 1 atm by applying the Langevin piston. Hydrogens were constrained with SHAKE,[31] allowing an integration step of 2 fs. Lookup tables[38] for the calculation of pairwise non-bonded interactions (van der Waals and Coulomb) were used to increase efficiency.

3.6.3 Clustering of trajectories

Clustering was applied to the MD snapshots (saved every 20 ps) to obtain the most populated conformers for iterative restarting of implicit solvent MD simulations. The first nanosecond of every trajectory was discarded. Pairs of snapshots were compared using the positional root mean square deviation (RMSD) upon optimal structural overlap, and clustering was performed by the Leader algorithm as implemented in the trajectory analysis program Wordom.[39]

The conformations of contiguous repeat pairs were clustered as follows: the N-terminal cap and the first internal repeat (N-cap/R1); the last internal repeat and the C-terminal cap (R4/C-cap); and all the internal repeats (Rn/Rn+1). As the pairs R1/R2, R2/R3, and R3/R4 are topologically identical, the conformations of the internal repeat pairs (R1/R2, R2/R3, and R3/R4) were collected together to increase the statistics and generate a single model for the internal repeat pair. Structures were clustered using the RMSD of C_α atoms (except for the first two residues for the N-cap and the last residue of the C-cap) and $C\gamma$ atoms to account for

the side chain orientation in the hydrophobic core. We excluded Cγ atoms of lysine, glutamine, asparagine, glutamate, and arginine residues because they are usually exposed to the solvent. Based on visual inspection of the structural dispersion of the most populated clusters, we selected a cutoff for RMSD clustering of 1.5 Å. For each cluster found its representative was extracted as the structure with the lowest RMSD from all the other cluster members.

3.6.4 Trajectory analysis

RMSD and root mean square fluctuation (RMSF) were calculated using as reference structures, respectively, the starting structure used in the dynamics and the structures averaged over 2 ns trajectory segments.

The quasiharmonic entropy was computed from the covariance matrix of the atomic fluctuations[40] using the trajectory analysis program Wordom.[39] Global entropies, calculated on all Cα atoms, were normalized by the number of residues in order to compare models of different lengths (e.g. YM_4A and $Y\bar{M}_4A$ have 243 residues while their variants $Y_{II}M_4A_{II}$ and $Y_{II}\bar{M}_4A_{II}$ have 242 residues). Local entropies were calculated for a subset of atoms spanning individual repeat dimers (i.e., N-cap/R1, R1/R2, R2/R3, R3/R4, and R4/C-cap).

3.6.5 Model generation

The initial armadillo model was derived from three homology models built with Insight II (Accelrys Inc.) by mapping the YM_4A (KK type) sequence onto the crystallographic structure of three natural ArmRPs: yeast karyopherin (importin-α), mouse importin-α, and murine β-catenin (PDB accession codes: 1EE4, 1Q1T, 2BCT, respectively). A single implicit solvent MD simulation was run for each homology model, whereas for further generation models, six MD simulations were run (data not shown).

The optimization of the initial position of hydrogens and subsequent energy minimization were performed with the CHARMM PARAM19[32] united atom force field with distance-dependent dielectric function. Loops connecting α-helices were relaxed through four minimization cycles consisting of 100 iterations of steepest descent and 200 steps of conjugate gradient algorithms with gradually decreasing harmonic restraints on the C_α atoms of the helices (i.e., force constants of 10, 5, 1, and 0.1 kcal mol^{-1} $Å^{-2}$).

The system was further optimized using the implicit solvent model FACTS[29] without restraints by 100 steps of steepest descent and 200 iterations of conjugate gradient, followed by an adopted basis Newton-Raphson minimizer, until an energy gradient of 0.02 kcal mol^{-1} $Å^{-1}$ was reached.

3.6.6 Design and synthesis of DNA encoding designed ArmRPs, protein expression and purification

Individual modules for the KK-type were assembled from overlapping primers (**Supp. Table 1**) as described previously[12] and cloned into a vector. Subsequently, to form proteins with identical internal modules, the single modules were PCR-amplified from the vectors and assembled as described[12]. Point mutations at position 26 and/or 29 (KK, QK KQ and QQ)

were introduced into the M-type consensus using site-directed mutagenesis (QuikChange, Stratagene). The modules were then digested from the vector with the type IIS restriction enzymes *Bpi*I and *Bsa*I and directly ligated together with similarly assembled original Y and A caps as described previously.[12] *BamH*I and *Kpn*I restriction sites were used for insertion of the whole genes into the vector pPANK and the plasmids were sequenced. For a more detailed description of the cloning procedure see the Supplementary Methods.

3.6.7 Protein purification

All unlabeled ArmRP variants were expressed in *E. coli* XL1-blue, and purified as described previously.[12]. Proteins for NMR studies were produced in the *E. coli* strain M15 (Qiagen) additionally containing the plasmid pREP4 (encoding *lacI*). Cells were grown in minimal medium with ^{15}N- ammonium chloride as the sole nitrogen source. The medium was supplemented with trace metals, 150 µM thiamine and 30 µg/ml kanamycin and 100 µg/ml ampicillin. Expression and purification by IMAC and gel filtration were performed as described previously.[12] Protein size and purity were assessed by 15% SDS-PAGE, stained with Coomassie PhastGel Blue R-350 (GE Healthcare, Switzerland). The expected protein masses were confirmed by SDS-PAGE and mass spectroscopy. Elution fractions from IMAC were passed over a desalting column (PD-10, GE Healthcare) to remove imidazole from the elution buffer.

3.6.8 Circular dichroism spectroscopy

All CD measurements were performed on a Jasco J-810 spectropolarimeter (Jasco, Japan) using a 0.5 mm or 1 mm circular thermo cuvette. CD spectra were recorded from 190-250 nm with a data pitch of 1 nm, a scan speed of 20 nm/min, a response time of 4 sec and a band width of 1 nm. Each spectrum was recorded three times and averaged. Measurements were performed at room temperature unless stated differently. The CD signal was corrected by buffer subtraction and converted to mean residue ellipticity (MRE). Heat denaturation curves were obtained by measuring the CD signal at 222 nm with temperatures increasing from 20°C to 95°C (data pitch, 1 nm; heating rate, 1°C/min; response time, 10 s; bandwidth, 1 nm). GdnHCl-induced denaturation measurements were performed after overnight incubation at 20°C with increasing concentrations of GdnHCl (99.5% purity, Fluka) in phosphate-buffered saline (pH 7.4).

3.6.9 ANS fluorescence spectroscopy

The fluorophore 1-anilino-naphthalene-8-sulfonate (ANS) binds to exposed hydrophobic patches or pockets in proteins, thereby increasing its fluorescence intensity. The measurements were performed at 20°C by adding ANS (final concentration 100 µM) to 10 µM of purified protein in 20 mM Tris·HCl, 50 mM NaCl, pH 8.0. The fluorescence signal was recorded using a PTI QM-2000-7 fluorimeter (Photon Technology International, USA). The emission spectrum from 400-650 nm (1 nm/s) was recorded with an excitation wavelength of 350 nm. For each sample, three spectra were recorded and averaged.

3.6.10 Size exclusion chromatography and multi-angle light scattering

The mass and oligomeric state of selected ArmRP was determined using a liquid chromatography system (Agilent LC1100), Agilent Technologies, Santa Clara, CA) coupled to an Optilab rEX refractometer and a miniDAWN three-angle light-scattering detector (both Wyatt Technology, Santa Barbara, CA). For protein separation a 24 ml Superdex 200 10/30 column (GE Healthcare Biosciences, Pittsburg, PA) was run at 0.5 ml/min in PBS. Typically, 50 µl of solution containing 30 µM protein was injected. Analysis of the data was performed using the ASTRA software (version 5.2.3.15; Wyatt Technology).

3.6.11 NMR Spectroscopy

Buffers used for NMR measurements of the internal repeat module optimization (KK- to QQ-type) contained 20 mM deuterated Tris·HCl, 30 mM NaCl, and the pH was adjusted to pH values of pH 8-11 using NaOH. All cap variants were analyzed in PBS buffer containing 150 mM NaCl and 50 mM sodium phosphate at pH 7.4. Proteins were concentrated to 0.5-1.0 mM for NMR measurements.

Proton-nitrogen correlation maps were derived from [^{15}N,^{1}H]-HSQC experiments[41] utilizing pulsed-field gradients for coherence selection and quadrature detection[42] and incorporating the sensitivity enhancement element of Rance and Palmer.[41, 42] All experiments were recorded on a Bruker AV-700 MHz spectrometer equipped with a triple-resonance cryoprobe at 310 K. Spectra were processed and analyzed in the spectrometer software TOPSPIN 2.1 and calibrated relative to the proton water resonance at 4.63 ppm, from which the ^{15}N scale was calculated indirectly using the conversion factor of 0.10132900.

Acknowledgement: We acknowledge financial support by the Swiss National Science foundation (SINERGIA credit No 122686).

3.7 References

1. Binz H K, Amstutz P, Plückthun A (2005) Engineering novel binding proteins from nonimmunoglobulin domains. Nat. Biotechnol. **23**: 1257-1268.
2. Boersma Y L, Plückthun A (2011) DARPins and other repeat protein scaffolds: advances in engineering and applications. Curr. Opin. Biotechnol. **22**: 849-857.
3. Lofblom J, Frejd F Y, Ståhl S (2011) Non-immunoglobulin based protein scaffolds. Curr. Opin. Biotechnol. **22**: 843-848.
4. Clonis Y D (2006) Affinity chromatography matures as bioinformatic and combinatorial tools develop. J. Chromatogr. A **1101**: 1-24.
5. Spisak S, Guttman A (2009) Biomedical applications of protein microarrays. Curr. Med. Chem. **16**: 2806-2815.
6. Andrade M A, Petosa C, O'Donoghue S I, Muller C W Bork P (2001) Comparison of ARM and HEAT protein repeats. J. Mol. Biol. **309**: 1-18.
7. Hatzfeld M (1999) The armadillo family of structural proteins. Int. Rev. Cytol. **186**: 179-224.
8. Marfori M, Mynott A, Ellis J J, Mehdi A M, Saunders N F, Curmi P M, Forwood J K, Boden M, Kobe B (2011) Molecular basis for specificity of nuclear import and prediction of nuclear localization. Biochim. Biophys. Acta **1813**: 1562-1577.
9. Tewari R, Bailes E, Bunting K A, Coates J C (2010) Armadillo-repeat protein functions: questions for little creatures. Trends Cell. Biol. **20**: 470-481.
10. Xu W, Kimelman D (2007) Mechanistic insights from structural studies of beta-catenin and its binding partners. J. Cell Sci. **120**: 3337-3344.
11. Cortajarena A L, Regan L (2006) Ligand binding by TPR domains. Protein Sci. **15**: 1193-1198.
12. Parmeggiani F, Pellarin R, Larsen A P, Varadamsetty G, Stumpp M T, Zerbe O, Caflisch A Plückthun A (2008) Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. J. Mol. Biol. **376**: 1282-1304.
13. Conti E, Uy M, Leighton L, Blobel G, Kuriyan J (1998) Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. Cell **94**: 193-204.
14. Conti E, Kuriyan J (2000) Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. Structure **8**: 329-338.
15. Huber A H, Weis W I (2001) The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. Cell **105**: 391-402.
16. Perrimon N, Mahowald A P (1987) Multiple functions of segment polarity genes in Drosophila. Dev. Biol. **119**: 587-600.
17. Wieschaus E, Riggleman R (1987) Autonomous requirements for the segment polarity gene armadillo during Drosophila embryogenesis. Cell **49**: 177-184.

18. MacDonald B T, Tamai K, He X (2009) Wnt/beta-catenin signaling: components, mechanisms, and diseases. Dev. Cell **17**: 9-26.
19. Mason D A, Stage D E, Goldfarb D S (2009) Evolution of the metazoan-specific importin alpha gene family. J. Mol. Evol. **68**: 351-365.
20. Moroianu J, Blobel G, Radu A (1996) Nuclear protein import: Ran-GTP dissociates the karyopherin alphabeta heterodimer by displacing alpha from an overlapping binding site on beta. Proc. Natl. Acad. Sci. U S A **93**: 7059-7062.
21. Interlandi G, Wetzel S K, Settanni G, Plückthun A, Caflisch A (2008) Characterization and further stabilization of designed ankyrin repeat proteins by combining molecular dynamics simulations and experiments. J. Mol. Biol. **373**: 837-854.
22. Slavik J (1982) Anilinonaphthalene sulfonate as a probe of membrane composition and function. Biochim. Biophys. Acta **694**: 1-25.
23. Montelione G T, Arrowsmith C, Girvin M E, Kennedy M A, Markley J L, Powers R, Prestegard J H, Szyperski T (2009) Unique opportunities for NMR methods in structural genomics. J. Struct. Funct. Genomics **10**: 101-106.
24. Kramer M A, Wetzel S K, Plückthun A, Mittl P R, Grütter M G (2010) Structural determinants for improved stability of designed ankyrin repeat proteins with a redesigned C-capping module. J. Mol. Biol. **404**: 381-391.
25. Wetzel S K, Ewald C, Settanni G, Jurt S, Plückthun A, Zerbe O (2010) Residue-resolved stability of full-consensus ankyrin repeat proteins probed by NMR. J. Mol. Biol. **402**: 241-258.
26. Zimm B H, Bragg J K (1959) Theory of the phase transition between helix and random coil polypeptide chains. J. Chem. Phys. **31**: 526-535.
27. Aksel T, Barrick D (2009) Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. Methods Enzymol. **455**: 95-125.
28. Brooks B R, Brooks III C L, Mackerell Jr A D, Nilsson L, Petrella R J, Roux B, Won Y, Archontis G, Bartels C, Boresch S (2009) CHARMM: the biomolecular simulation program. J. Comp. Chem. **30**: 1545-1614.
29. Haberthür U, Caflisch A (2008) FACTS: Fast analytical continuum treatment of solvation. J. Comp. Chem. **29**: 701-715.
30. Brooks B R, Bruccoleri R E, Olafson B D, Swaminathan S, Karplus M (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. **4**: 187-217.
31. Ryckaert J P, Ciccotti G, Berendsen H J C (1977) Numerical-integration of cartesian equations of motion of a system with constraints - molecular-dynamics of n-alkanes. J. Comput. Phys. **23**: 327-341.
32. MacKerell Jr A D, Bashford D, Bellott M, Dunbrack Jr R L, Evanseck J D, Field M J, Fischer S, Gao J, Guo H, Ha S (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B **102**: 3586-3616.
33. Mackerell A D, Jr., Feig M Brooks C L, 3rd (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J. Comp. Chem. **25**: 1400-1415.

34. Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W, Klein M L (1983) Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. **79**: 926-935.
35. Darden T, York D Pedersen L (1993) Particle Mesh Ewald - An n.log(n) method for Ewald sums in large systems. J. Chem. Phys. **98**: 10089-10092.
36. Hoover W G (1985) Canonical dynamics: Equilibrium phase-space distributions. Phys. Rev. A. **31**: 1695-1697.
37. Nosé S (1984) A unified formulation of the constant temperature molecular-dynamics methods. J Chem Phys **81**: 511-519.
38. Nilsson L (2009) Efficient table lookup without inverse square roots for calculation of pair-wise atomic interactions in classical simulations. J. Comput. Chem. **30**: 1490-1498.
39. Seeber M, Felline A, Raimondi F, Muff S, Friedman R, Rao F, Caflisch A, Fanelli F (2011) Wordom: A user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. J. Comput. Chem. **32**: 1183-1194.
40. Andricioaei I, Karplus M (2001) On the calculation of entropy from covariance matrices of the atomic fluctuations. J. Chem. Phys. **115**: 6289-6292.
41. Bodenhausen G, Ruben D J (1980) Natural Abundance Nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem. Phys. Lett. **69**: 185-189.
42. Keeler J, Clowes R T, Davis A L, Laue E D (1994) Pulsed-field gradients: theory and practice. Meth. Enzymol. **239**: 145-207.
43. Conti E, Kuriyan J (2000) Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. Structure **8**: 329-338.

3.8 Supplementary Material

Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy

Pietro Alfarano[1¥], Gautham Varadamsetty[1¥], Christina Ewald[2], Fabio Parmeggiani[1], Riccardo Pellarin[1], Oliver Zerbe[2], Andreas Plückthun[1*], Amedeo Caflisch[1*]

[1] Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

[2] Department of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

3.8.1 Supplementary Materials and Methods

3.8.1.2 Design and synthesis of DNA encoding designed ArmRPs

The cloning, expression, and purification of designed ArmRPs was carried out essentially as described previously.[1] All primers were synthesized and purified by Microsynth GmbH (Balgach, Switzerland) (**Table S1**). The melting temperature (T_m) was calculated as described (http://www.stratagene.com/manuals/200519.pdf). First, single modules were assembled. In case of the KK-module assembly, in a first step partially overlapping primer pairs (1)-(2), (3)-(4) and (5)-(6) (**Table S1**) were annealed and the double strand was completed by PCR. Then, 2 µl from these PCR reaction mixtures were combined as templates for a second assembly reaction in the presence of oligonucleotides (1) and (6). All the oligonucleotides were used at final concentrations of 1 µM. The annealing temperature was 50°C for the first and second reaction. Thirty PCR cycles were performed with an extension time of 30 s. The same procedure was applied for the internal and other capping repeats. Four oligonucleotides were used for the N-terminal capping repeats.

BamHI and KpnI restriction sites were used for direct insertion of modules into plasmid pQE30 and the modules were sequenced. The single modules were PCR amplified from the vectors, using external primers pQE_f_1 and pQE_r_1. Neighboring modules were digested with restriction enzymes BpiI and BsaI and directly ligated together. The genes coding for the whole proteins were assembled by stepwise ligation of the internal and capping modules. BamHI and KpnI restriction sites were used for insertion of whole genes into the vector pPANK. Proper assembly of constructs was validated by DNA sequencing.

Q and K mutations were introduced by QuikChange mutagenesis, carried out in 50 µl with 50-100 ng template, 0.4-2 µM primer pair, 200 µM dNTPs and 2 u of *Pfu* DNA polymerase (Stratagene, CA). The reaction was initiated by pre-heating to 94°C for 3 min; followed by 18 cycles of 94°C for 1 min, 52°C for 1 min and 68°C for 15 min, followed by incubation at 68°C for 1 h. Chemically competent E. coli XL1-Blue cells were transformed with an aliquot of 5 µl of DpnI-digested PCR products (to remove the original plasmid

serving as PCR template) and inoculated on Luria–Bertani (LB) agar plates containing 100 µg/ml ampicillin. To further remove any residual wild-type plasmid from potentially doubly transformed cells a total of 5 colonies for each mutant were picked and their plasmids were isolated by mini-prep, digested, and the insert was ligated into fresh vector. The positive mutants were verified by sequencing analysis.

For the generation of N-cap and C-cap mutants, the pQE30-based plasmid pPANK[1] containing the designed ArmRP YM_4A gene fragment was used as the PCR mutagenesis template for introducing N- and C-cap mutations.

To assemble the entire YM_4A protein, the modules were digested with the type IIS restriction enzymes BpiI and BsaI and directly ligated together and this step was repeated to result in the quadruple M_4 pieces that were then ligated to the corresponding caps. BamHI and KpnI restriction sites were used for insertion of the whole genes into the vector pPANK and the plasmids were sequenced.

3.8.1.3 **Reference**

1. Parmeggiani F, Pellarin R, Larsen A P, Varadamsetty G, Stumpp M T, Zerbe O, Caflisch A Plückthun A (2008) Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. J. Mol. Biol. **376**: 1282-1304.

Table S3.1 List of oligonucleotides used for generating point mutants

Oligonucleotides		
name	**sequence 5'-3' direction**	**description (for=forward, rev=reverse)**
Y_{II}_f	GCTACCCGTAAATTCTCTCAGATCCTGTCTGATGGTAACGAACAAATC	for primer for introducing three mutations in N-cap
Y_{II} _r	CCATCAGACAGGATCTGAGAGAATTTACGGGTAGCAGACAGCTGTTCC	rev primer for introducing three mutations in N-cap
C Q240M_f	CTGGAGAAGATGTTCTCCCACTAATGAGGTACCCCGG	for primer for introducing Met at position 240 in C-cap
C Q240M_r	GTGGGAGAACATCTTCTCCAGAGCTTCCTGAGCTTCTTTC	rev primer for introducing Met at position 240 in C-cap
C Q240L_f	CTGGAGAAGCTGTTCTCCCACTAATGAGGTACCCCGG	for primer for introducing Leu at position 240 in C-cap
C Q240L_r	GTGGGAGAACAGCTTCTCCAGAGCTTCCTGAGCTTCTTTC	rev primer for introducing Leu at position 240 in C-cap
C Q240M,F24Q_f	CTGGAGAAGATGCAGTCCCACTAATGAGGTACCCCGGGTC	for primer for introducing Met at position 240 and Gln at position 241 in C-cap
C Q240M,F241Q_r	CATTAGTGGGACTGCATCTTCTCCAGAGCTTCCTGAGCTTCTTTC	rev primer for introducing Met at position 240 and Gln at position 241 C-cap
A_{II} cap_f	CTGGAGAAGCTGCAGTCCCACTAATGAGGTACCCCGGGTCG	for primer for introducing Leu at position 240 and Gln at position 241 in C-cap
A_{II} cap_r	CATTAGTGGGACTGCAGCTTCTCCAGAGCTTCCTGAGCTTCCTGAGCTTCTTTC	for primer for introducing Leu at position 240 and Gln at position 241 in C-cap
Cons_1_for (1)	CCAGGGATCCTAGGAAGACCTTGGTAACGAACAAATCC	for assembly consensus M module of KK type and amplification
2A_rev (2)	AGCCGGCAGAGCACCAGCATCGATAACAGCTTGGATTTGTTCGTTACCAAGG	rev assembly consensus M module of KK type
3A_for (3)	GGTGCTCTGCCGGCTCTGGTTCAACTGCTGTCCTCTCCGAACG	for assembly consensus M module of KK type
4L_rev (4)	CCACAGAGCTTCTTTCAGGATCTTCTCGTTCGGAGAGGACAGC	rev assembly consensus M module of KK type
5L-I_for (5)	CTGAAAGAAGCTCTGTGGGCTCTGTCTAACATCGCTTCTGGTGGTTGAG	for assembly consensus M module of KK type
6I_rev (6)	TTCCTGGTACCCTAAGGTCTCAACCACCAGAAGCGAT	rev assembly consensus M module of KK type and amplification
KQ_for	CCGAACGAGAAGATCCTGCAAGAAGCTCTGTGGGC	for mutation KK->KQ
KQ_rev	GCCCACAGAGCTTCTTGCAGGATCTTCTCGTTCGG	rev mutation KK->KQ
QK_for	CCTCTCCGAACGAGCAGATCCTGAAAGAAGC	for mutation KK->QK
QK_rev	GCTTCTTTCAGGATCTGCTCGTTCGGAGAGG	rev mutation KK->QK
QQ_for	CCTCTCCGAACGAGCAGATCCTGCAAGAAGCTCTGTGGGC	for mutation KK->QQ
QQ_rev	GCCCACAGAGCTTCTTGCAGGATCTGCTCGTTCGGAGAGG	rev mutation KK->QQ
pQE_f_1	CGGATAACAATTTCACACAG	forward primer for pQE vectors
pQE_r_1	GTTCTGAGGTCATTACTG	reverse primer for pQE vectors

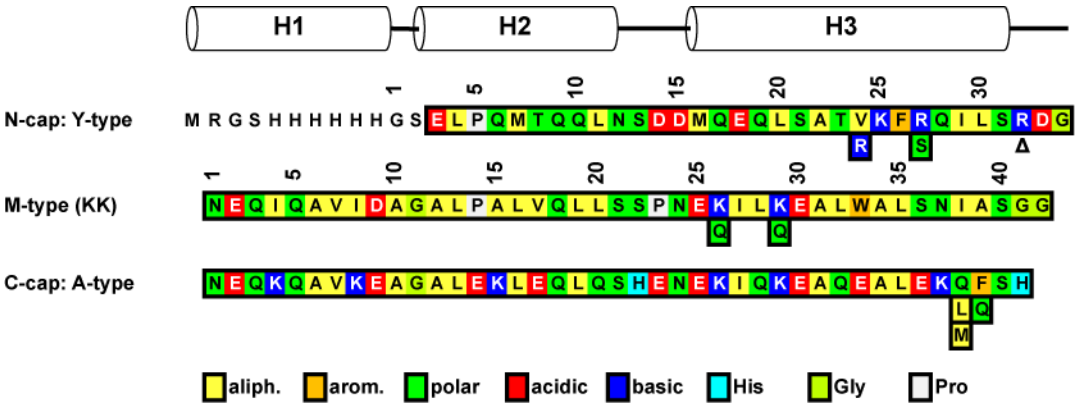


Figure S3.1 Sequence of the armadillo repeats and mutants studied. In the proteins studied experimentally, the N-cap is preceded by the sequence MRGSHHHHHHGS for purification and detection. Experimentally, YM_3A and YM_4A molecules have been tested. KK refers to the internal repeat as shown, QK with only a substitution at position 26, KQ with only a substitution at position 29, and QQ with both, as shown underneath. Y_{II} refers to an N-cap carrying all three mutations indicated underneath, A_{II} refers to a C-cap with the mutations L and Q indicated underneath

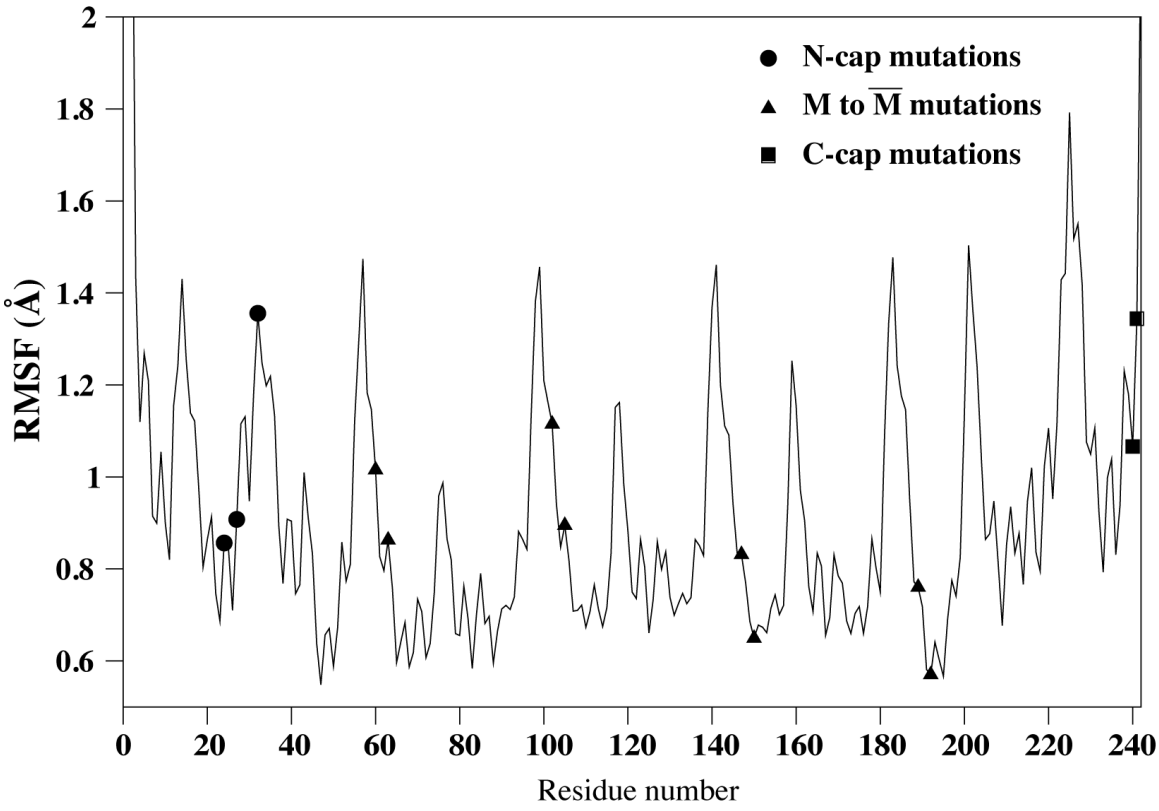


Figure S3.2 RMSF plot of Cα atoms of YM_4A in the implicit solvent simulations. The circles show the position of the residues suggested for mutations at the N-cap: Val-24, Arg-27 and Arg-32. The triangles show the internal repeat positions Lys-26 and Lys-29; and the squares the residues suggested for mutations at the C-cap: Gln-240 and Phe-241

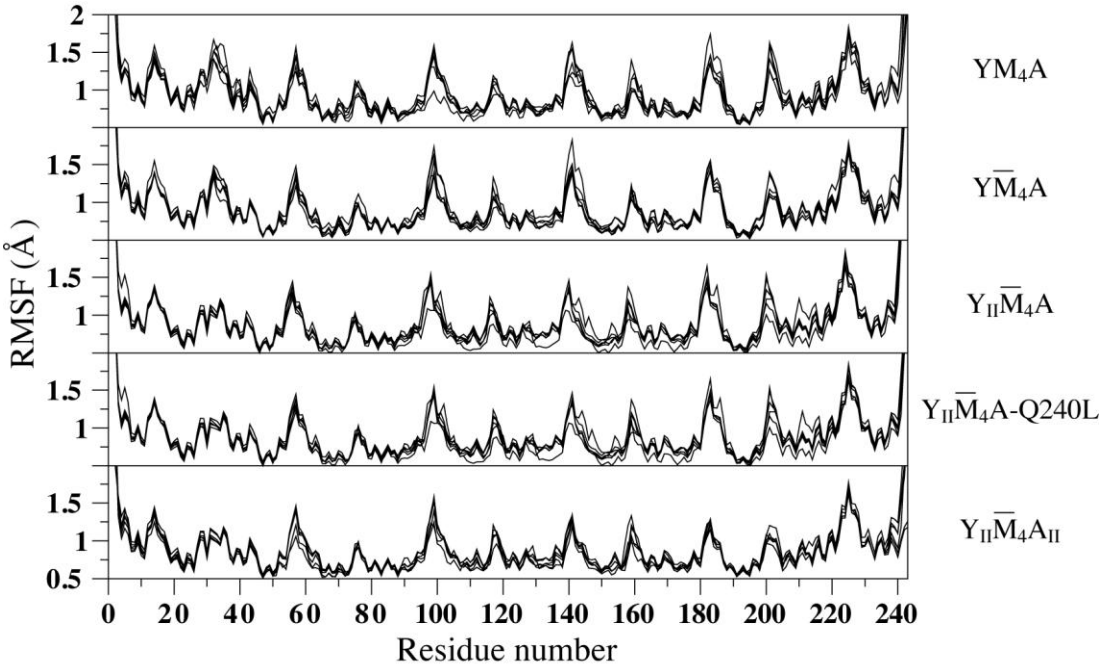


Figure S3.3 RMSF plot superposition of all implicit solvent simulations

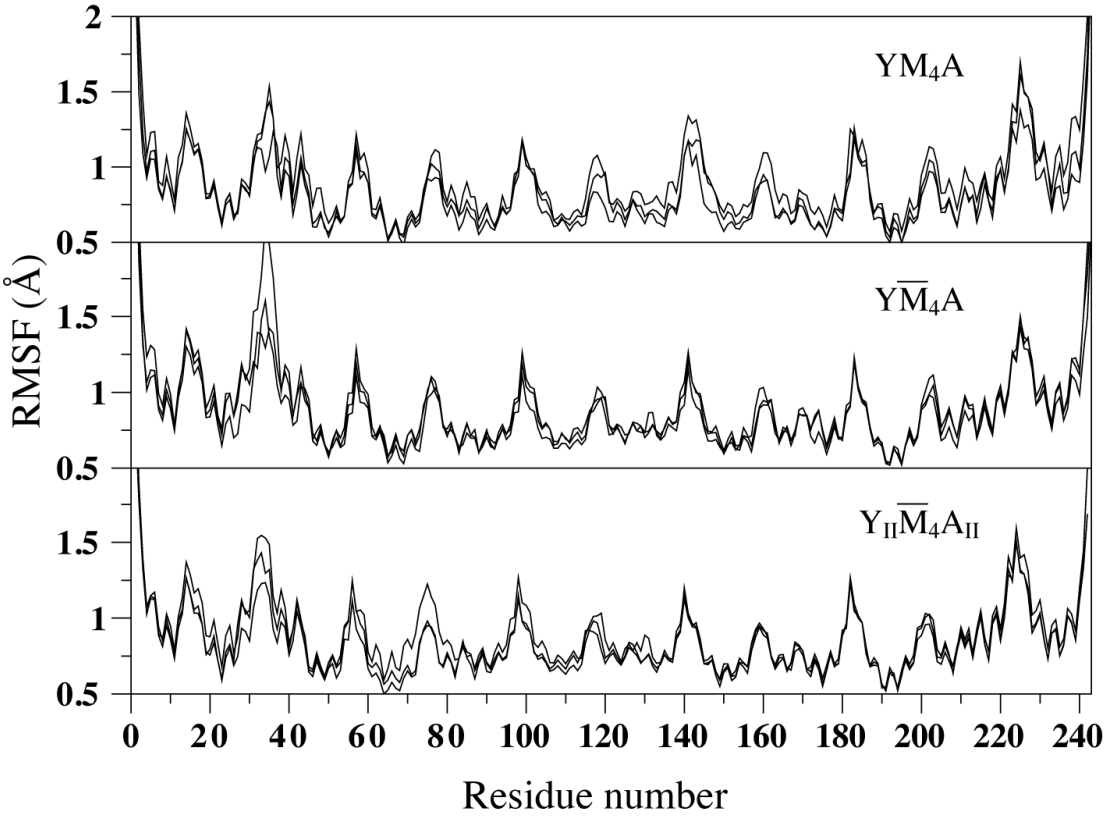


Figure S3.4 RMSF plot superposition of all explicit solvent simulations

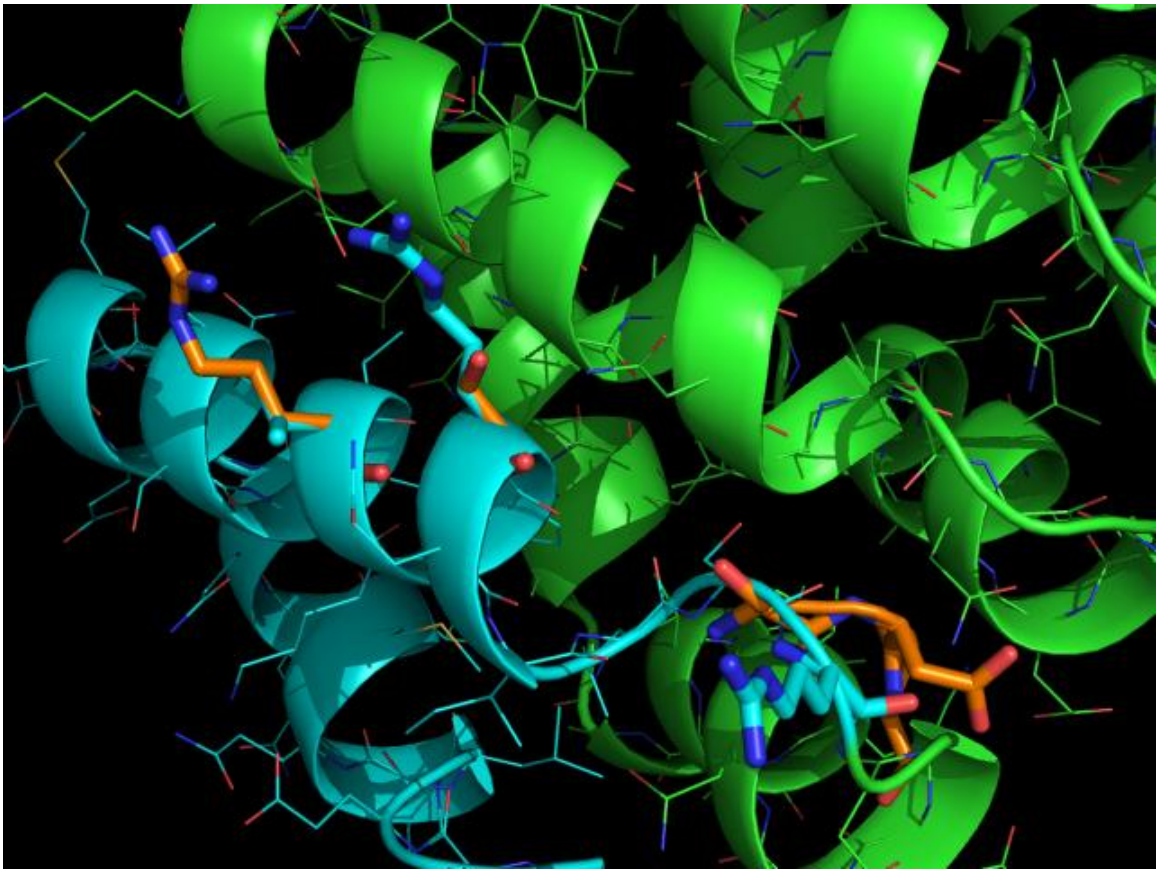


Figure S3.5 Mutations introduced into the N_{II}-cap compared to wild-type. Wild-type residues (V24, R27 and R32) are in depicted in cyan while the three mutations introduced in the N_{II}-cap (R24, S27 and ΔR32 (deletion of arginine) are shown in orange. Helices colored in cyan refer to the N_{II}-cap while helices in green to the internal repeats

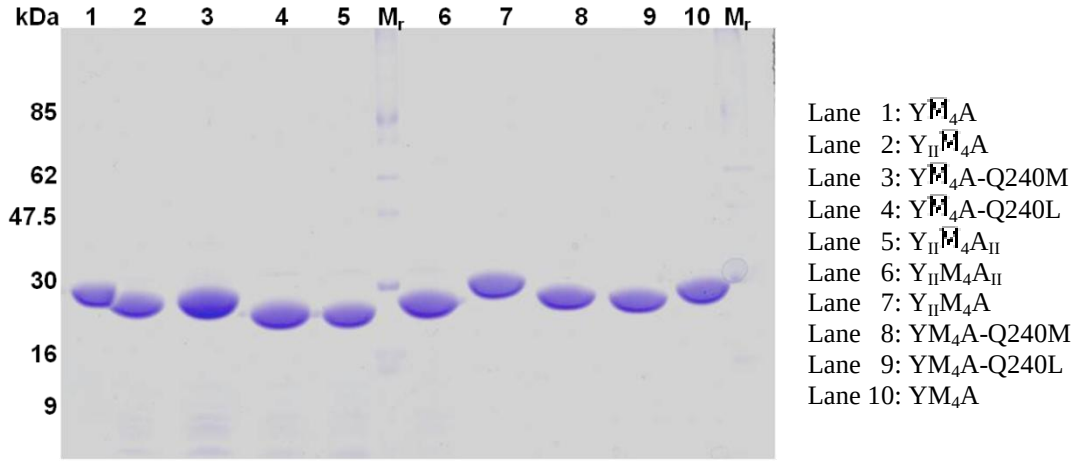


Figure S3.6 Single-step IMAC purification of YM/$\bar{M}_4$A N-and C-cap variants. The expected size is approximately 27 kDa. Proteins were analyzed by SDS-PAGE (15%). The size marker is indicated in kDa

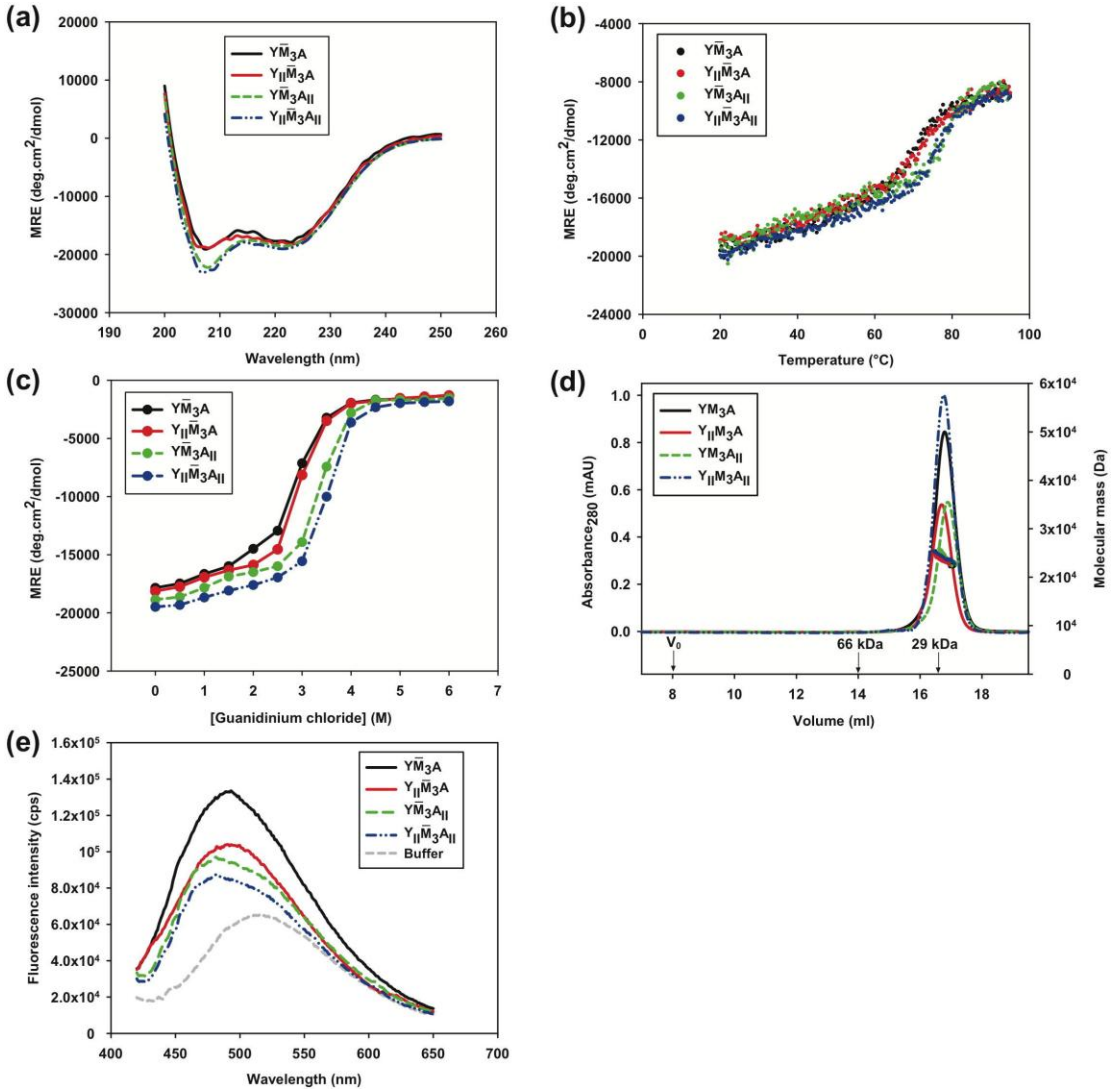


Figure S3.7 Biophysical characterization of designed ArmRPs $YM̄_3A$ and its cap variants ($Y_{II}M̄_3A$, $YM̄_3A_{II}$ and $Y_{II}M̄_3A_{II}$). (a) **CD spectra,** (b) **thermal denaturation curves and** (c) **GdnCl-induced denaturation curves. The denaturation experiments were followed by CD. The values of MRE at 222 nm are reported.** (d) **SEC and MALS of designed ArmRPs. The absorbance at 280 nm from SEC is shown on the left y-axis, the calculated MW from MALS on the right y-axis. V_0 indicates the void volume of the column. Bovine serum albumin (MW 66 kDa), and carbonic anhydrase (MW 29 kDa) were used as molecular weight markers, and the corresponding elution volumes are indicated by the arrows.** (e) **ANS binding. The values without buffer subtractions are shown. The protein concentration used was 10 μM for a,b,c,e and 30 μM for d**

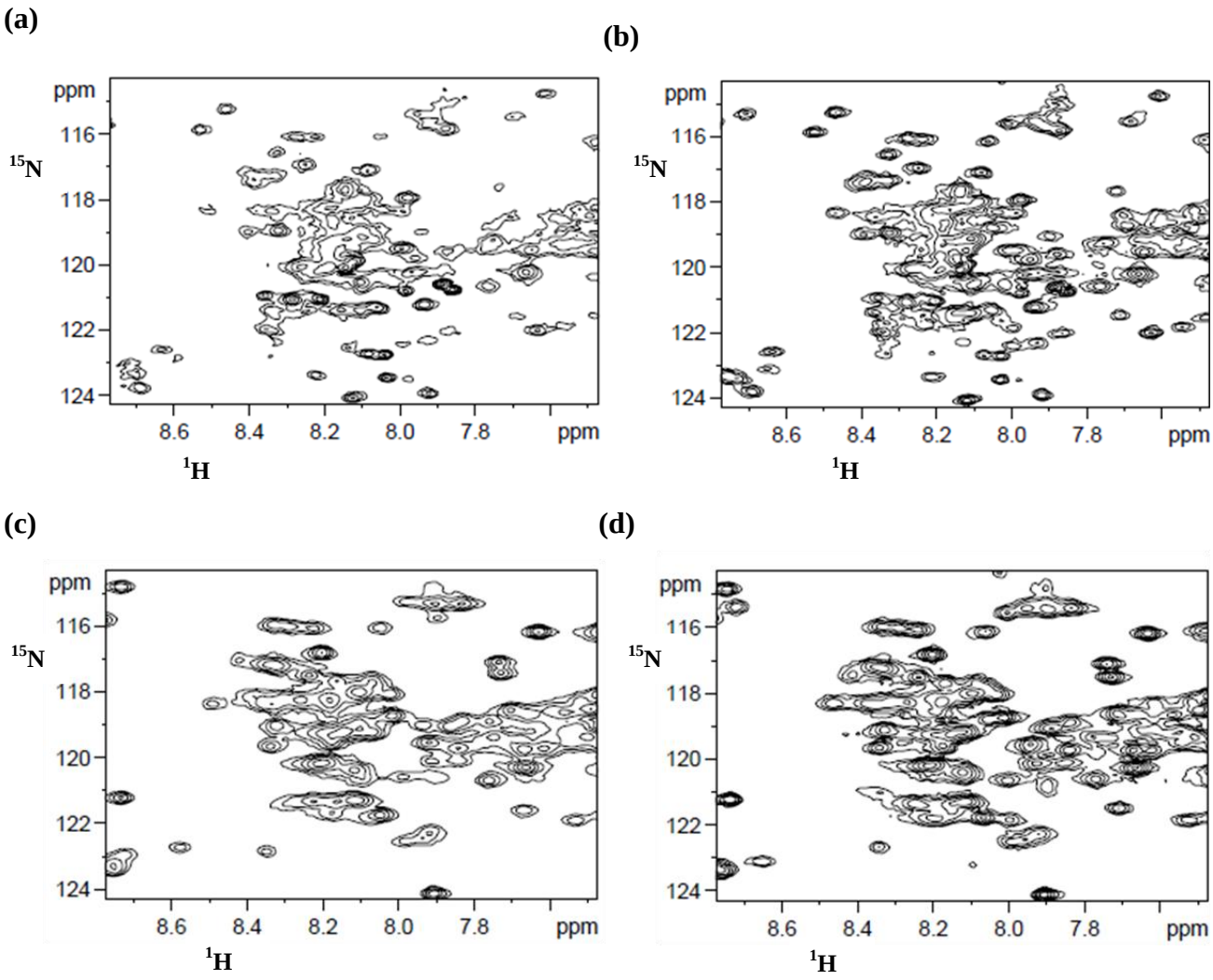


Figure S3.8 [^{15}N, ^{1}H]-HSQC spectra of designed ArmRP Y$\bar{M}_4$A (a) **and its cap variants $Y_{II}\bar{M}_4A$** (b), **$Y\bar{M}_4A_{II}$** (c) **and $Y_{II}\bar{M}_4A_{II}$** (d) **at pH 7.4. All spectra were recorded at a temperature of 310 K in 50 mM phosphate buffer, and 150 mM NaCl. The protein concentration was 0.5 mM. All proteins are of the QQ ($\bar{M}$) – type**

4 Spontaneous Self-Assembly of Fragments of Engineered Armadillo Repeat Proteins into Folded Proteins

Randall P. Watson[1], Fabian Bumbak[1], Christina Ewald[1], Christian Reichen[2], Andreas Plückthun[2]*, Oliver Zerbe[1]*

[1] Institute of of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

[2] Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

*corresponding authors:

Email address of corresponding authors: oliver.zerbe@oci.uzh.ch; plueckthun@bioc.uzh.ch

4.1 Introduction

The characteristic feature of repeat proteins is the packing of multiple identical amino acid stretches[1,2] into folded modules which rigidly associate into stable proteins. Typically, a short motif of 20-50 amino acids forms such a folded module, and these are repeated with highly similar sequences within the same protein. Within a repeat protein the modules fold into nearly identical structures. In the repeat proteins the stacked structural modules form an extended domain with a continuous surface. Since its composition can often be varied while maintaining the structure it is often employed in forming specific interactions, often with a high affinity[3]. Examples of such proteins are the ankyrin repeat proteins[4], HEAT repeats[1] and the armadillo repeat proteins[5-8]. Many of these protein are involved in cell signaling or transport[9].

Armadillo repeat proteins (ArmRPs) bind peptides in an extended form, and thus it is the amino acid sequence of the peptide rather than its tertiary structure that is recognized[8,10]. Furthermore, there is an approximate correspondence between the modules and two consecutive side chains being recognized. Accordingly, armadillo repeat proteins make particularly attractive scaffolds for protein engineering and biotechnological applications[3]. For these reasons, Parmeggiani *et al.* [11] developed designed repeat proteins based on combined sequences of the natural armadillo repeat proteins of β-catenins and importin-α. In these proteins, a special N-terminal repeat (or N-cap) and a special C-terminal repeat (C-cap) are used to flank the internal repeats. Recently, first crystal structures have been determined, verifying the consensus design[12].

Nonetheless, NMR would complement crystallography in many parts of the design cycle. Yet, assignment of chemical shifts of these proteins by NMR is very challenging due to the repetitive nature of their sequence[13]. To facilitate this process we attempted segmental labeling[14] using a split intein[15,16] to aid in deconvolution of the intrinsically complex degenerate spectra. We observed that, when the repeat protein was expressed as two separate fragments with their intein ligation motifs present, the fragments showed an affinity for each other, although no peptide bond was formed. Removal of the split intein motifs resulted in the same observation, indicating that the interaction was not mediated by the split intein. This indicated the formation of a stable, non-covalent complex. We present structural, biophysical and thermodynamic data to characterize this interaction. Furthermore we analyse this interaction with reference to the structure of the complete protein, and demonstrate that the same interfacial contacts are made, indicating that the interaction occurs in a highly similar if

not identical manner. This finding not only has implications for future applications but may also shed light on the evolution of repeat proteins.

4.2 Results and Discussion

As a model system we investigated a consensus armadillo repeat protein consisting of three identical internal repeats, flanked by N- and C-terminal capping repeats. The fragments were recombinantly expressed and purified as two fragments from separate *E. coli* cultures as described in the Supplementary Materials. Most detailed investigations were carried out with an N-terminal fragment consisting of an N-terminal capping repeat (N-cap) and two internal repeats (hereafter called YM_2, the Y denoting the yeast origin of the N-cap, and the M denoting the Molecular Dynamics origin of the internal repeats[11]) and a C-terminal fragment (termed MA) consisting of one internal repeat and a C-terminal capping repeat (A for artificial)[11]. Amino acid sequences of the fragments are given in Figure 4.1.

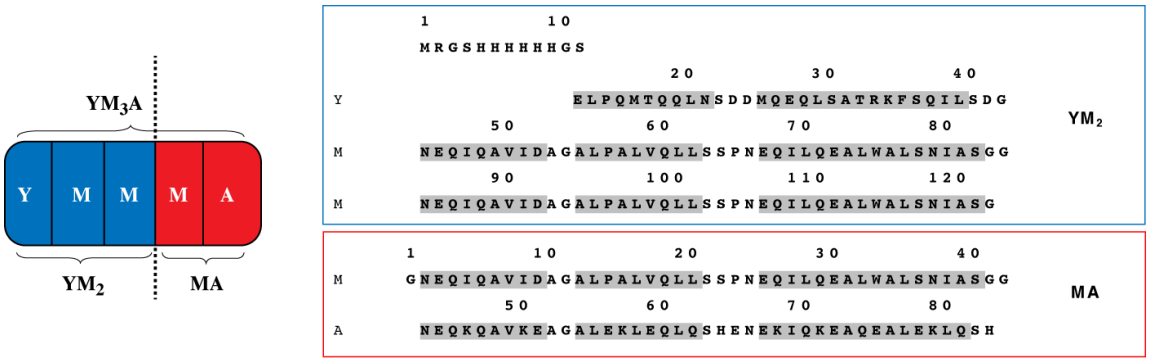


Figure 4.1 Amino acid sequences of the two fragments investigated in this study

4.2.1 Properties of the fragments

The [^{15}N, ^{1}H]-HSQC spectrum of the C-terminal MA fragment displays good signal dispersion and narrow peaks, both features indicative of a well-folded protein (Figure 4.2A). In contrast, spectra of YM_2 are essentially completely devoid of peaks from backbone resonances, a behavior typically associated with a protein lacking well defined tertiary structure such as a molten globule (Figure 4.2C)[17]. This result is remarkable, as in general the stability of repeat proteins, including ArmRP, increases with length[11,18,19], and YM_2 contains one repeat more than MA.

Upon mixing ^{15}N-labeled YM_2 with a slight excess of unlabeled MA, the [^{15}N, ^{1}H]-HSQC spectrum for YM_2 resembles data from of a well behaved soluble protein (Figure 4.2D),

indicating that adding the C-terminal fragment has enabled YM_2 to become folded. In the reverse experiment, ^{15}N-labeled MA was mixed with unlabeled YM_2. Compared to the [^{15}N, ^{1}H]-HSQC of the uncomplexed MA fragment many of the resonances shift to new positions, indicating a structural change associated with formation of a complex with YM_2 (Figure 4.2B and Figure S 4.10).

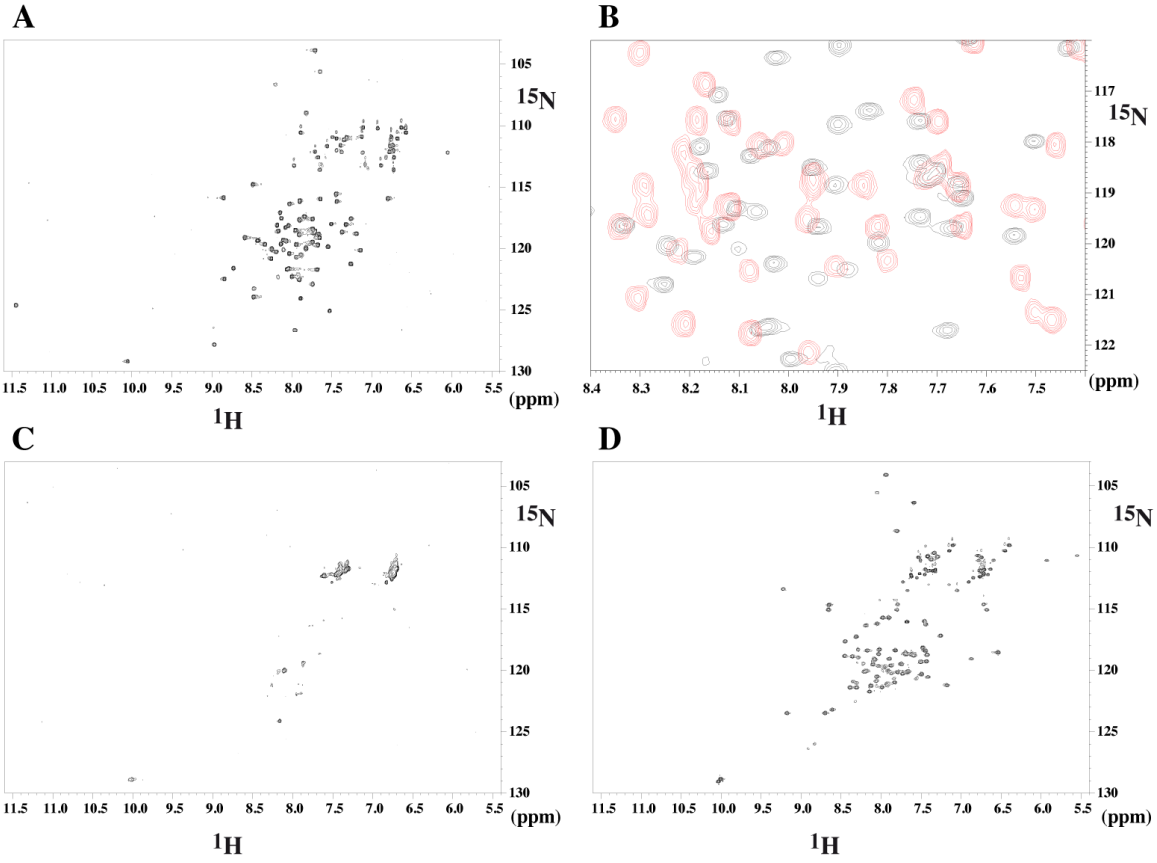


Figure 4.2 [^{15}N, ^{1}H]-HSQC spectra of: A, ^{15}N MA; B, Expansion of the spectrum of ^{15}N MA complexed with unlabeled YM_2 (red) at a 1:1.2 molar ratio, superimposed with the spectrum in absence of YM_2 (black); C, ^{15}N labeled YM_2 alone; D, ^{15}N labeled YM_2 complexed with unlabeled MA at 1:1.2 molar ratio. All spectra recorded at 310 K in 50 mM sodium phosphate buffer with 150 mM NaCl, 2 % glycerol, 0.02 % NaN_3, 10 % D_2O, pH 7.4.

Far-UV CD spectra (Supp. Mat. Figure S4.3) of the individual fragments display features typical of helical proteins. Predictions of the helical proportion using the program K2D[20] for MA and YM_2 are 76% and 40%, respectively. MA is clearly in a predominantly helical state, but CD data of YM_2 surprisingly also indicate a high degree of helical secondary structure. Melting curves of MA followed by CD show a marked sigmoidal transition at ~62°C, characteristic of cooperative folding (Supp. Mat. Figure S4.4). The melting curves for YM_2,

however, are essentially straight lines with a poorly defined transition. Near-UV CD spectra (not shown) of both YM_2 and MA gave no interpretable signal, expected from the lack of buried aromatic residues in the cores — the few aromatics present being solvent exposed in the crystal structure of the unsplit protein [12].

The absence of peaks in the [^{15}N, ^{1}H]-HSQC spectrum of uncomplexed YM_2 (Figure 4.2C) is most likely due to insufficient packing of side chains. To exclude the possibility that the lack of peaks is due to formation of large oligomers we have used size-exclusion chromatography (SEC) (Supp. Mat. Figure S4.2). The position of the elution peak of YM_2 (nominal MW 12.2 kDa) showed a marked correlation between concentration and oligomeric state. An injection of a sample at 60 µM resulted in a single narrow and symmetric peak, indicating a unique species with an apparent molecular weight of 36 kDa, whereas multiple peaks were observed at concentrations of 300 µM. [^{15}N,^{1}H]-HSQC spectra of ^{15}N YM_2 collected at 60 µM resulted in spectra similar to those collected at higher concentrations, again suggesting that the lack of peaks is due to the absence of tertiary structure and not oligomerization.

It is likely, therefore, given the apparent molecular size determined from SEC and the monomeric like behavior at these lower concentrations, that the YM_2 fragment is in a molten globule or pre-molten globule like state. As a result of this poor packing a significant amount of exposed hydrophobic surface renders the fragment susceptible to limited aggregation. This aggregation is apparently easily reversed by the addition of the complexing MA partner fragment. In contrast, the NMR spectrum of the MA fragment does not change between 200 and 800 µM.

4.2.2 Properties of the complex

NMR experiments described above indicate that the complementary fragments form a stable complex in solution. We then utilized isothermal titration calorimetry (ITC), with buffers and temperature identical to those used in the NMR experiments, to determine the thermodynamic properties of the interaction. A titration experiment at 305 K in which MA was added to YM_2 yielded a K_d of ~126 ± 5 nM with corresponding ΔH of -78.2 kJ mol^{-1} and $T\Delta S$ of -37.9 kJ mol^{-1} (Figure 4.3).

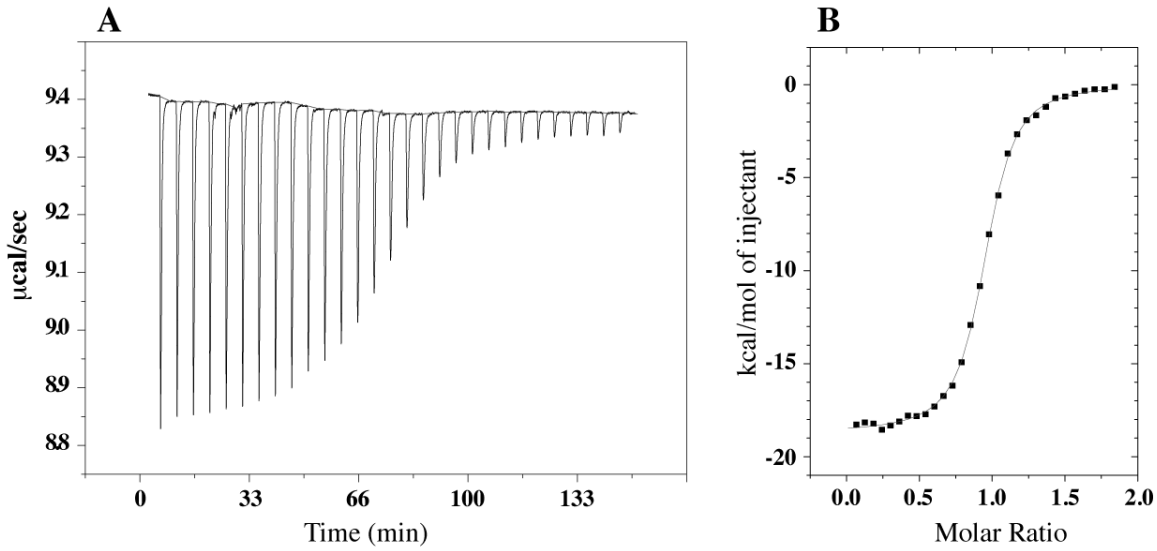


Figure 4.3 Isothermal titration calorimetry isotherm and curve fitting for the YM_2/MA interaction.

We suggest that the measured K_d is heavily influenced by the entropy penalty of refolding the N-terminal fragment, and to a lesser degree by the restriction of N-terminal conformational exchange in the C-terminal fragment (*vide infra*), even though the overall entropy term is still favorable because hydrophobic surface is buried. It is likely that if the N-terminal fragment could be engineered to adopt a more structured form before binding the K_d would be even lower.

To complement the biophysical data we have used solution NMR to determine the conformations of the fragments both isolated and when in complex with each other. The solution structure of uncomplexed MA was solved using a $^{15}N,^{13}C$-labeled protein and 2D and 3D NMR (Figure 4.4A, Figure 4.4B, Figure S4.5, Figure S4.8 and Table S4.6). The structure was calculated using 788 NOE-derived distance restraints. The root-mean-square-deviation (RMSD) amongst the 20 lowest energy conformers is 0.80±0.21 Å for backbone atoms of residues 14 to 82. The structure is depicted in Figure 4.4A and shows remarkable structural similarity with the corresponding region from the crystal structure of the full-length protein (PDB # 4DBA). However, signals from the 14 N-terminal residues, corresponding to the 1st helix of the internal consensus repeat, were absent in the spectra – most likely due to conformational exchange. The RMSD of the mean coordinates of MA to the crystal structure when superimposed for residues 14 to 82 is 2.44 Å for backbone atoms and 2.81 Å for all heavy atoms.

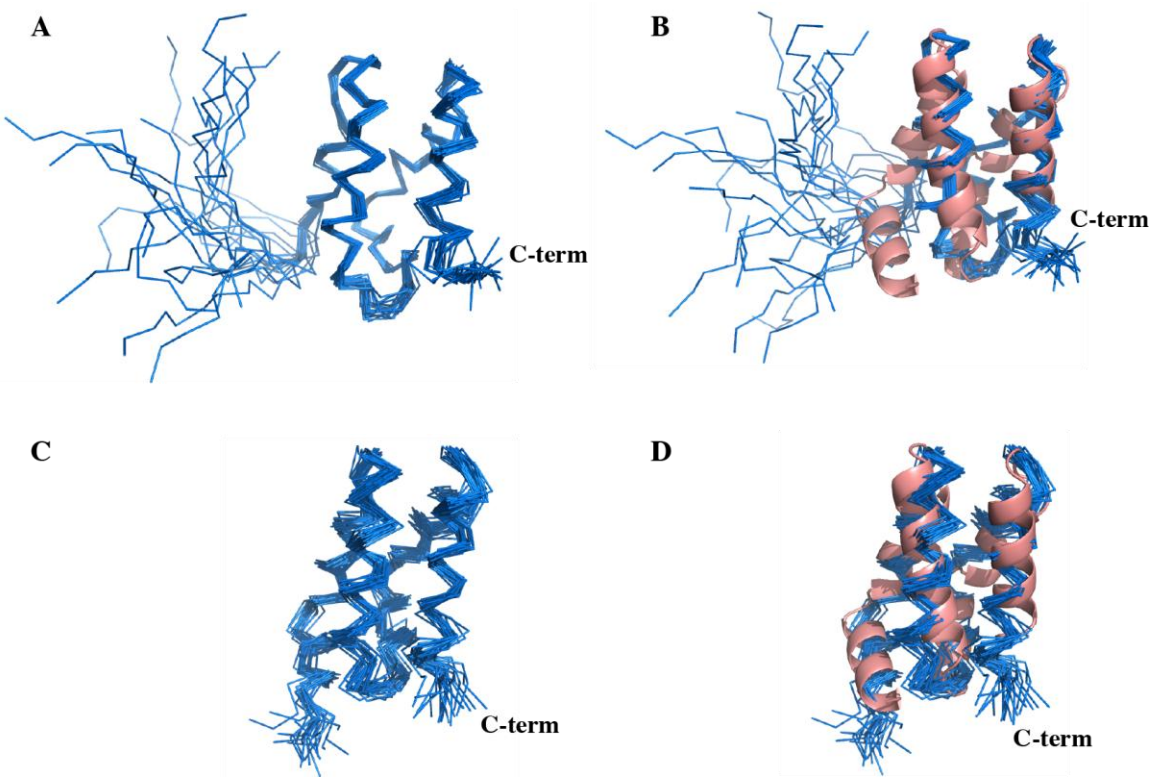


Figure 4.4 A, Solution structure of uncomplexed MA displaying the ensemble of the 20 lowest-energy conformers. B, Solution structure of uncomplexed MA superimposed with the corresponding region from the crystal structure of the entire protein, YM_3A. C, Solution structure of MA complexed with YM_2. D, Ensemble of the 20 lowest-energy structures superimposed over the crystal structure of YM_3A.

In addition, the solution structure of MA complexed with unlabeled YM_2 was solved (Figures Figure 4.4C, Figure 4.4D, Figure S4.6, Figure S4.9 and Table S4.7). In this structure only 3 of the initial 14 unassigned N-terminal residues remained unassigned. The root-mean-square-deviation (RMSD) amongst the 20 lowest energy conformers is 1.17 Å for backbone atoms of residues 4 to 82. The structure reveals that the newly assigned N-terminal residues are now present in the form of a stable helix and are no longer in slow conformational exchange. The overall structure closely resembles the corresponding region in the crystal structure, the backbone atoms of residues 4 to 82 aligning with an RMSD of 2.48 Å.

Finally we investigated the N-terminal fragment in presence of the complementing C-terminal fragment. Unlike for the smaller MA fragment, assignment of the YM_2 spectra proved more problematic due to the presence of two stretches of largely identical amino acid sequence in YM_2. To date, backbone assignments have been accomplished for the C-terminal (M) repeat and the helices 2 and 3 in the 1st internal (M) repeat of the YM_2 fragment (Supp. Mat. Figure S4.7). A considerable proportion of the YM_2 fragment is in conformational exchange; specifically the N-terminal capping repeat (Y) and the 1st helix of the 1st internal

repeat (M). Predictions of secondary structure from the assigned backbone chemical shifts using the program TALOS[21] indicate for the assigned residues of the internal repeats good coincidence between secondary structure and the crystal structure (Supp. Mat. Figure S4.11). Again, the N-terminal capping repeat is in conformational exchange on an intermediate time scale.

Assignment of the majority of the side-chain resonances in the 2nd internal repeat in YM_2 has been accomplished and has allowed the use of ^{13}C edited, ^{15}N,^{13}C filtered NOESY experiments[22] to observe intermolecular NOEs between side-chain protons of the labeled YM_2 and the unlabeled ^{12}C-H protons of the complexing MA species. In addition, these experiments have also been carried out with the reverse labeling scheme.

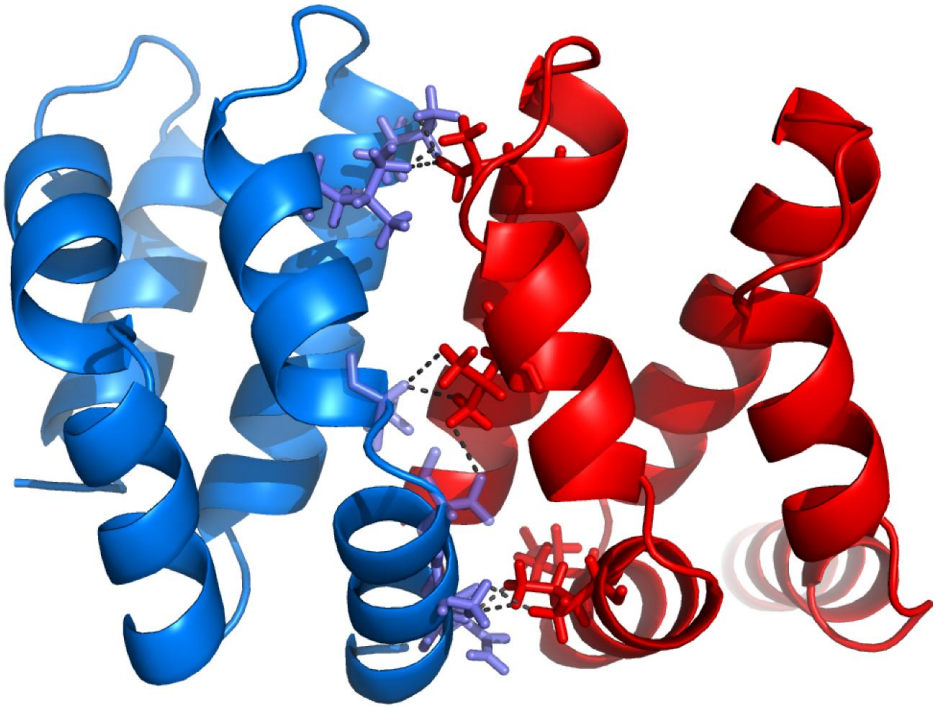


Figure 4.5 Image indicating MA (blue) and YM_2 (red) and a selection of observed inter-molecular NOEs, establishing the relative orientations of the two fragments. Image based on the known crystal structure of the complete YM_3A molecule (PDB# 4DBA). Note that the structure has been rotated about the vertical axis by 180 degrees such that the N-terminus points to the right.

From a total of 73 intermolecular NOEs originating from YM_2, 4 could be unambiguously attributed to protons from MA within a 3 Å distance (Figure 4.5). When increasing the range to 6 (or 9) Å to account for side-chain rotamers different from those in the crystal structure of the entire protein, an additional 17(or 30) NOEs can be unambiguously assigned. Moreover, 10 (or 14) more NOEs from ambiguous YM_2 proton assignments can be matched to MA

protons. These ambiguities arise from the difficulties caused by the identical repeat units. It is likely that the few remaining NOEs, which are currently unaccounted for, will be assigned during further refinement, in particular when MA side-chain conformations are computed from data that additionally include restraints from YM_2. Despite the fact that not all details of the interaction are determined yet, the presence of these inter-molecular NOEs indicates that the complex is formed such that the two fragments are in a native-like orientation, where many native interactions seem to occur that are made through adaptations of side-chain conformations.

Assuming a native-like complex interface, the buried interface is ~ 920 $Å^2$ and according to our NMR structure ~ 2400 $Å^2$ of the uncomplexed MA surface is buried during complex formation, explaining the rather good affinity despite the fact that energy is lost upon folding both the N- and C- termini. Importantly, most of the previously solvent exposed residues of MA are thereby shielded from solvent in the complex.

4.3 Conclusions

We have demonstrated in this work that a consensus-designed Armadillo repeat protein, when split into two fragments, is capable of regaining the structure of the parent protein at the interface through formation of a non-covalent complex. In this work the C-terminal MA fragment was structured to a large degree and served as a template onto which the N-terminal YM_2 fragment bound in a coupled folding-binding event with remarkably high affinity. Often, the truncation or splitting of a folded globular protein exposes the hydrophobic core to the solvent and usually results in severe aggregation and precipitation. This includes other repeat proteins, such as Designed Ankyrin Repeat proteins, where fragments missing one capping repeat have an aggregation tendency [23].

This feature of Armadillo Repeat proteins lends itself to exploitation in the design of binding proteins from fragments, and a rather rapid engineering by exchanging non-covalent protein fragments. Nonetheless, it will have to be investigated whether even more favorable assemblies can be found by using different breakage points at positions inside the repeats. We are presently investigating the scope of this result in a broader context.

Finally, we believe that this result may be important when considering the way in which these repeat proteins have evolved in nature. It is generally assumed that repeat proteins have arisen by gene duplication of the repeats [24], but a functional gain from a complex of assembled fragments may drive selection for longer repeat proteins. In the case of consensus

ArmRP, the solubility of the fragments is high enough that this ability can be directly shown. While many protein have been split into fragments that can reassemble[25], the ArmRP assembly may have lead to a rapid evolution of functional repeats that have allowed their widespread use in binding of extended peptides of different sequence, possibly including protein fragments as an evolutionary intermediate.

4.4 Acknowledgements

We acknowledge financial support by the SINERGIA program of the Swiss National Science Foundation (grant no 122686). We are indebted to Dr. I. Jeselarov for help with the ITC measurements and assistance in interpretation of the data.

4.5 References:

[1] M. A. Andrade, C. Perez-Iratxeta, C. P. Ponting, *J. Struct. Biol.* **2001**, *134*, 117-131.

[2] E. M. Marcotte, M. Pellegrini, T. O. Yeates, D. Eisenberg, *J. Mol. Biol.* **1999**, *293*, 151-160.

[3] Y. L. Boersma, A. Plückthun, *Curr. Opin. Biotechnol.* **2011**, *22*, 849-857.

[4] S. G. Sedgwick, S. J. Smerdon, *Trends Biochem. Sci.* **1999**, *24*, 311-316.

[5] M. Hatzfeld, *Int. Rev. Cytol.* **1999**, *186*, 179-224.

[6] M. Marfori, A. Mynott, J. J. Ellis, A. M. Mehdi, N. F. Saunders, P. M. Curmi, J. K. Forwood, M. Boden, B. Kobe, *Biochim. Biophys. Acta* **2011**, *1813*, 1562-1577.

[7] R. Tewari, E. Bailes, K. A. Bunting, J. C. Coates, *Trends Cell. Biol.* **2010**, *20*, 470-481.

[8] W. Xu, D. Kimelman, *J. Cell Sci.* **2007**, *120*, 3337-3344.

[9] B. T. MacDonald, K. Tamai, X. He, *Dev. Cell* **2009**, *17*, 9-26.

[10] A. H. Huber, W. I. Weis, *Cell* **2001**, *105*, 391-402.

[11] F. Parmeggiani, R. Pellarin, A. P. Larsen, G. Varadamsetty, M. T. Stumpp, O. Zerbe, A. Caflisch, A. Plückthun, *J. Mol. Biol.* **2008**, *376*, 1282-1304.

[12] C. Madhurantakam, G. Varadamsetty, M. G. Grütter, A. Plückthun, P. R. Mittl, *Protein Sci.* **2012**,

[13] S. K. Wetzel, C. Ewald, G. Settanni, S. Jurt, A. Plückthun, O. Zerbe, *J. Mol. Biol.* **2010**, *402*, 241-258.

[14] T. Yamazaki, T. Otomo, N. Oda, Y. Kyogoku, K. Uegaki, N. Ito, Y. Ishino, H. Nakamura, *J. Am. Chem. Soc.* **1998**, *120*, 5591-5592.

[15] C. Ludwig, D. Schwarzer, J. Zettler, D. Garbe, P. Janning, C. Czeslik, H. D. Mootz, *Biophys. J.* **2009**, *462*, 77-96.

[16] M. Muona, A. S. Aranko, V. Raulinaitis, H. Iwai, *Nat Protoc* **2010**, *5*, 574-587.

[17] H. J. Dyson, P. E. Wright, *Chemical Rev.* **2004**, *104*, 3607-3622.

[18] S. K. Wetzel, G. Settanni, M. Kenig, H. K. Binz, A. Plückthun, *J. Mol. Biol.* **2008**, *376*, 241-257.

[19] T. Kajander, A. L. Cortajarena, E. R. Main, S. G. Mochrie, L. Regan, *J. Am. Chem. Soc.* **2005**, *127*, 10188-10190.

[20] M. A. Andrade, P. Chacón, J. J. Merelo, F. Morán, *Protein Eng* **1993**, *6*, 383-390.

[21] Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, *J. Biomol. NMR* **2009**, *44*, 213-223.

[22] G. Otting, K. Wüthrich, *Q. Rev. Biophys.* **1990**, *23*, 39-96.

[23] G. Interlandi, S. K. Wetzel, G. Settanni, A. Plückthun, A. Caflisch, *J. Mol. Biol.* **2008**, *373*, 837-854.

[24] J. Lee, M. Blaber, *Proc. Natl. Acad. Sci. U S A* **2011**, *108*, 126-130.

[25] S. S. Shekhawat, I. Ghosh, *Curr. Opin. Chem. Biol.* **2011**, *15*, 789-797.

4.6 Supplementary Materials

Spontaneous Self-Assembly of Fragments of Engineered Armadillo Repeat Proteins into Folded Proteins"

by R. P. Watson et al.

4.6.1 Methods and Materials

4.6.1.1 Cloning of designed Armadillo Repeat Protein Fragments.

Oligonucleotides were purchased form Microsynth AG (Balgach, Switzerland). A complete list of all oligonucleotides used is given in Table S4.1.

For ligase-independent cloning (LIC), based on the method of Aslanidis *et al.*[1], vector pLIC_CR containing an N-terminal $MRGSH_6$-tag followed by a rTEV recognition site and a *Sac*B gene as additional selection marker was constructed from vector eLIC_043 (Neldner et al., unpublished), by *Eco*RI and *Sfi*I digestion and PCR amplified primers LIC_148_for and LIC_148_rev.

All fragments (YM, YM_2, M_2A and MA) and full length construct (YM_3A) were amplified from pQE-N1M3C1[2] with the following primers: LIC_N_for and LIC_M_rev for YM and YM_2; LIC_M_for and LIC_C_rev for M_2A and MA; LIC_N_for and LIC_C_rev for YM_3A. Amplification was performed with an annealing temperature of 65 °C and the corresponding bands where gel purified. 100 ng of *Bsa*I digested vector and fragment, both T4-DNA-polymerase treated for LIC cloning in the presence of dCTP and dGTP respectively, were mixed and incubated for 10 min at room temperature before *E. coli* XL1-blue cells were transformed and grown on plates containing 100 µg/ml AMP and 7 % sucrose.

Table S4.1 Sequences of Oligonucleotide Primers

Name	Sequence 5'-3' direction
LIC_148_for	CACAGAATTCATTAAAGAGGAGAAATTAAC
LIC_148_rev	TGGGCCGGCTGGGCCGGTCTCGAAAATATAAATTTTCGGATCCGTGATGGTGATGGTGATGC
LIC_M_for	GAAAATTTATATTTTCAGGGGAACGAACAAATCCAAGCTGTTATCGATGC
LIC_C_rev	AGATGAGAGTAAGGCTATCATTAGTGGGACTGCAGCTTCTCCAGAGC
LIC_N_for	GAAAATTTATATTTTCAGGGGGAACTGCCGCAGATGACCCAGCAGCTGAACTCC
LIC_M_rev	AGATGAGAGTAAGGCTATCATTAACCAGAAGCGATGTTAGACAGAGCCCACAGAGC

4.6.1.2 **Protein expression**

YM_2 and MA were expressed in *E. coli* M15 (pREP4) (Qiagen) transformed with expression plasmids based on the vector pLIC and encoding the mature YM2 or MA proteins. The crude expression products are shown in Figure S4.1 and summarized in Table S4.2.

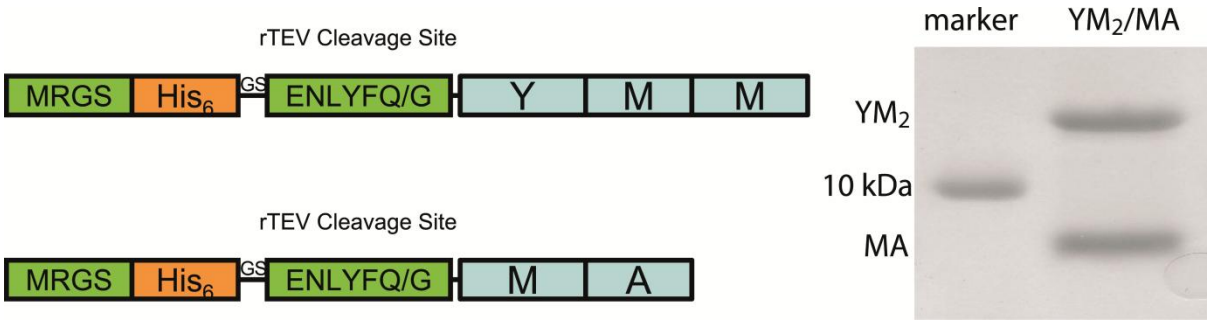


Figure S4.1 Schematic overview of expression products YM_2 (top left) and MA (bottom left) and 15% SDS-PAGE analysis of the rTEV treated, Ni-column purified products before further purification by SEC.

Table S4.2 Molecular weights of unlabeled fragments

Construct w/o isotopic labeling	MW [kDa]
YM_2 w/o His-tag	12.2
MA w/o His-tag	9.1
YM_2/MA complex both w/o His-tag	21.3

For unlabeled protein expression a scrape of colonies from a plate of *E. coli* M15 (pREP4) freshly transformed with the pLIC-based expression vector were cultured at 37 °C overnight in 5 mL LB medium supplemented with 100 mg/L ampicillin and 50 mg/L kanamycin. The overnight culture was added to 500 mL LB medium supplemented with antibiotics as described above. Cultures were grown in baffled 2 L Erlenmeyer flasks at 37 °C and 220 rpm. Cultures were induced at an OD_{600} of 0.5 with 1 mM IPTG and further incubated for 5-6 h. Cells were harvested by centrifugation at 9000 *g* for 20 min. Cell pellets were stored at -20 C.

For ^{15}N, ^{13}C labelled protein expression 5 mL of an LB overnight culture, generated as previously described, were centrifuged at 2000 *g* for 10 min to remove unlabeled media, minimize cellular degradation products and to remove β-lactamase. The recovered cell mass was resuspended in 0.5 L minimal media prepared as described in Table S4.3. All isotopically labelled compounds were acquired from Cambridge Isotopes, UK.

Table S4.3 Composition of minimal medium

^{15}N labeling	mM	**^{15}N,^{13}C labeling**	mM
K_2HPO_4	22.97	K_2HPO_4	22.97
KH_2PO_4	29.39	KH_2PO_4	29.39
$Na_2HPO_4 \cdot 2H_2O$	5.80	$Na_2HPO_4 \cdot 2H_2O$	5.80
$^{15}NH_4Cl$	9.35	$^{15}NH_4Cl$	9.35
D-(+)-glucose	25.23	^{13}C-D-(+)-glucose	10.09
Thiamine	0.15	Thiamine	0.15
$MgSO_4$	2.00	$MgSO_4$	2.00

Table S4.4 Composition of trace metal solution

1000 × trace metal stock	g/L	mM
$FeSO_4 \cdot 7H_2O$	4	14.39
$CaCl_2 \cdot 2H_2O$	4	27.21
$AlCl_3 \cdot 6H_20$	1	3.93
$MnSO_4 \cdot H_2O$	1	5.92
$CoCl_2 \cdot 6H_20$	0.4	1.59
$ZnSO_4 \cdot 7H_2O$	0.2	0.70
$CuCl_2 \cdot 2H_2O$	0.1	0.68
H_3BO_4	0.1	1.28

Minimal media were supplemented with 50 mg/L ampicillin and 12.5 mg/L kanamycin and 1× trace metal solution (Table S4.4). Cultures in 2 L baffled Erlenmeyer flasks were incubated at 37° C and 220 rpm. At an OD_{600} of 0.5, cultures were induced with 1 mM IPTG. For ^{15}N labeling, expression cultures were harvested after 6 h, for ^{15}N,^{13}C labeling, expression cultures were harvested after 12 h.

4.6.1.3 **Protein Purification**

Cell pellets were thawed at RT and 25 mL TBS_{500} (50 mM Tris·HCl pH 8.0, 500 mM NaCl, 5 % glycerol) was added. Cells were lysed by sonication on ice. Sonication was carried out at 27 % amplitude (Branson Digital Sonifier, Model 250 with microtip) with a pulse length of 10 s and a rest period of 15 s for a total of 10 min. Cellular debris was pelleted by centrifugation at 17,000 *g* in SS34 tubes for 20 min at 4 °C. Supernatants were filtered through 0.2 µm cellulose acetate syringe filters (Sarstedt) and kept on ice. Cell lysates were

loaded onto 2×1 mL (concatenated) or 5 mL Hi-Trap Ni columns (Pharmacia) pre-equilibrated with TBS_{500}. Columns were washed with 10 column volumes (CV) TBS_{500}. For MA, the columns were washed further with 5 CV TBS_{500}, 5 mM imidazole. MA was eluted from the column using a single elution step of 5 CV TBS_{500}, 50 mM imidazole. For YM_2, the columns were washed with TBS_{500}, 50 mM imidazole, and YM_2 was eluted using a gradient from 50-500 mM imidazole over 15 CV. YM_2 eluted between 120 mM and 160 mM imidazole.

Eluted proteins were treated with rTEV protease (to remove the His_6 affinity purification tag) at a molar ratio of 1:30 enzyme:protein for MA and 1:5 for YM_2. Treatment was at RT during which proteins were dialyzed against 200 × volume of PBS_{150} (50 mM phosphate buffer pH 7.4, 150 mM NaCl, 2 % glycerol). The digestion progress was monitored by SDS-PAGE. After digestion was complete, the solution was filtered (0.2 µm cellulose acetate syringe filters, Sarstedt) to remove precipitated rTEV (a small amount of the rTEV precipitated during the dialysis process) and the solution was passed through a 2×1 mL pre-equilibrated Hi-Trap Ni column (Pharmacia) to remove cleaved His_6 peptide and the also His_6-tagged rTEV, and any impurities which co-eluted with YM_2 or MA in the previous imidazole elution.

4.6.1.4 **Size Exclusion Chromatography**

All rTEV digested fragments as well as the YM_2/MA complex were further purified by preparative SEC to obtain samples for NMR spectroscopy. MA was concentrated to a final volume of ~5 mL containing 100 to 300 µM protein and injected onto a preparative S75 SEC column (S75 16/60 HiLoad, GE Healthcare) at 1 mL/min in PBS_{150} pH 7.4, 2 %. Fractions of 1 mL were collected over the peak centered at 78 mL (Figure S4.2).

For experiments, where YM_2 was required alone, it was purified by SEC in the same manner. However, the elution profile of YM_2 showed that the protein exhibited a strong concentration-dependent oligomerization/aggregation effect, with a series of poorly resolved peaks being eluted from the void volume (47 mL) with a final major peak at 58 mL. Because of this, for experiments where uncomplexed YM_2 was required, it was only purified by SEC at <100 µM, and used after SEC without any further concentrating. For the purification of YM_2 to be used in complex with MA, the complex was formed first and the mixture then purified by SEC. For NMR experiments requiring complete complexation of isotopically labelled YM_2, a greater than equimolar amount (1.5:1) of MA to YM_2 was used. The mixtures

were concentrated to ~5 mL for injection onto the column. The complex was eluted at 68 mL and excess uncomplexed MA eluted at 78 mL. For NMR experiments requiring the complete complexation of labelled YM_2 the excess MA fractions were added back to the fractions of the complex. For NMR experiments requiring complete complexation of the labelled MA, a small amount of YM_2 was added to the complex peak fractions estimated to enable a slight excess of YM_2, and the whole solution was concentrated to a level suitable for NMR measurements (0.25-1 mM).

Analytical SEC was carried out using an S200 5/150 GL (Pharmacia) column on an ÄKTA HPLC system. PBS_{150} (pH 7.4, 2 % glycerol) as described above was used, the column was run at 0.3 mL/min and injections contained 50 µL.

The SEC columns were calibrated using the following mixtures of proteins: 2 MDa Blue Dextran, 67 kDa albumin (17-0442A, Pharmacia LMW std), 44.3 kDa ovalbumin (A-5378, Sigma, Chicken egg), 25 kDa chymotrypsinogen (17-04542B, Pharmacia LMW std), 13.7 kDa ribonuclease A (R-5503, Sigma, from Bovine Pancreas). The apparent deviations of the uncomplemented fragments from the expected size (Table S4.5) are caused by their non-compact fold. This deviation is considerably reduced in the complex which shows a smaller deviation from the globular standards, similar to full-length ArmRP constructs (Varadamsetty *et al.*, submitted). Armadillo repeat proteins have been shown to elute at higher apparent sizes, as expected from their elongated shape.

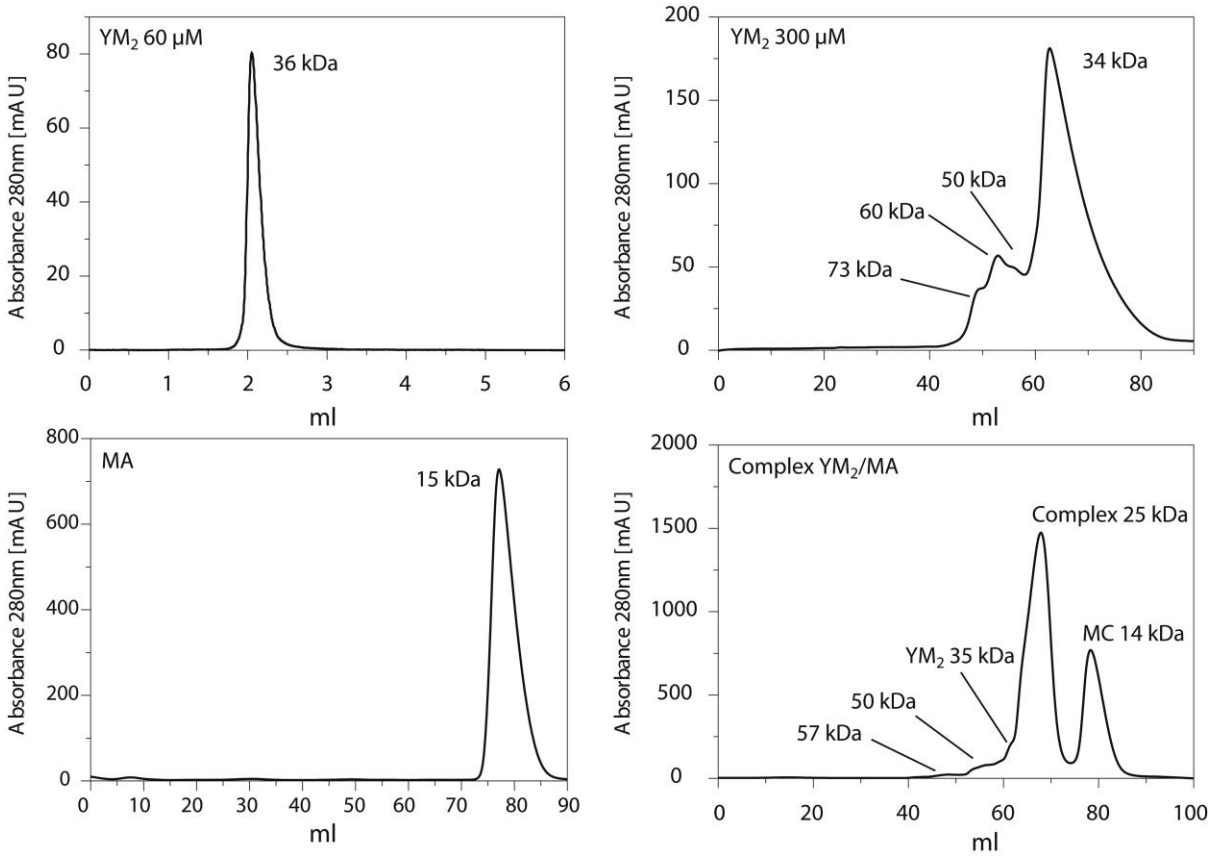


Figure S4.2 Preparative and analytical size exclusion analysis: YM_2 behaves as a monomer at low concentrations (top left, 60 µM) and shows concentration-dependent aggregation at higher concentrations (top right, 300 µM). MA elutes as one monomeric peak (bottom left). The complex of YM_2 and MA elutes as one peak (bottom right), uncomplexed MA and YM_2 fragments are observed at their respective positions in the elution profile.

Table S4.5 Comparison of sequence based molecular weight and SEC-observed size

Construct	MW [kDa]	Apparent size in SEC [kDa]	Column
YM_2 low conc. (60 µM)	12.2	36	S200 5/150 GL
YM_2 high conc. (300 µM)	12.2	34	S75 16/60 HiLoad
MA	9.1	15	S75 16/60 HiLoad
Complex YM_2/MA	21.3	25	S75 16/60 HiLoad

4.6.1.5 **Isothermal Titration Calorimetry**

Unlabeled YM_2 and MA were used for ITC experiments. Samples of YM_2 (20 µM 5 mL) and MA (70 µM, 8 mL) were separately dialyzed against 2× 2 L PBS_{150} pH 7.4, 2 % glycerol at room temperature, each buffer change lasting 12 h. YM_2 was diluted to 6.7 µM with the dialysis buffer, MA was used at 69.7 µM.

ITC was carried out using a VP-ITC (MicroCal Ltd.) instrument at 305 K. YM_2 was in the cell initially, and MA was added during the titration. The cell volume was 1.47 mL. 32 injections, each of 10 µL of MA, were made at 300 s intervals. Data were integrated and curve fitting carried out using Origin Software.

4.6.1.6 **Circular Dichroism**

Proteins obtained from the same preparations as used for the ITC measurements were used for CD measurements. Measurements were recorded using a JASCO J-715 instrument fitted with a JASCO PFD425S Peltier-controlled cuvette holder. A path length of 1 cm was used. Samples were diluted with MQ water to reduce detector overload. Final measurement concentrations were: 2 µM protein in 5 mM sodium phosphate buffer, pH 7.4, 15 mM NaCl, 0.2 % glycerol. Simple scans were measured at 100 nm/min in replicates of 5. A slit width of 1 nm was found to give adequate signal, and an integration time of 0.125 s was used. Melting curves were measured at 220 nm, with a heating rate of 1°C min^{-1}. Both YM_2 and MA displayed α-helical characteristics, however, YM_2 appears considerably less structured (Figure S4.3).

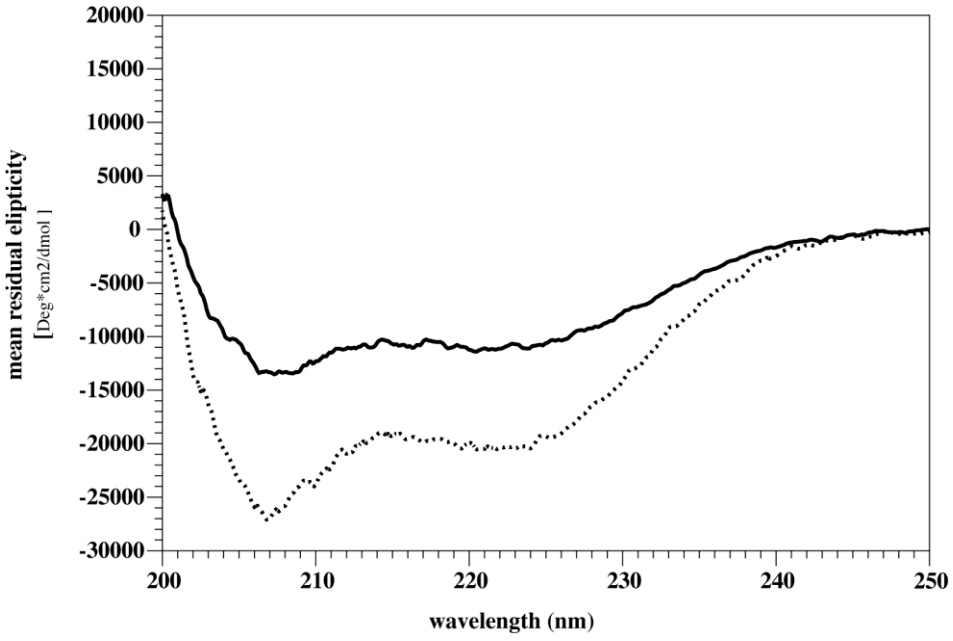


Figure S4.3 CD spectra of YM_2 (solid line) and MA (dotted line).

The melting point of MA was found to be 62 °C, whereas an exact melting temperature for YM_2 could not easily be determined due to the flat character of the curve without a true

transition point or plateau (Figure S4.4). At this low concentration the thermal denaturation of both YM_2 and MA was reversible (data not shown).

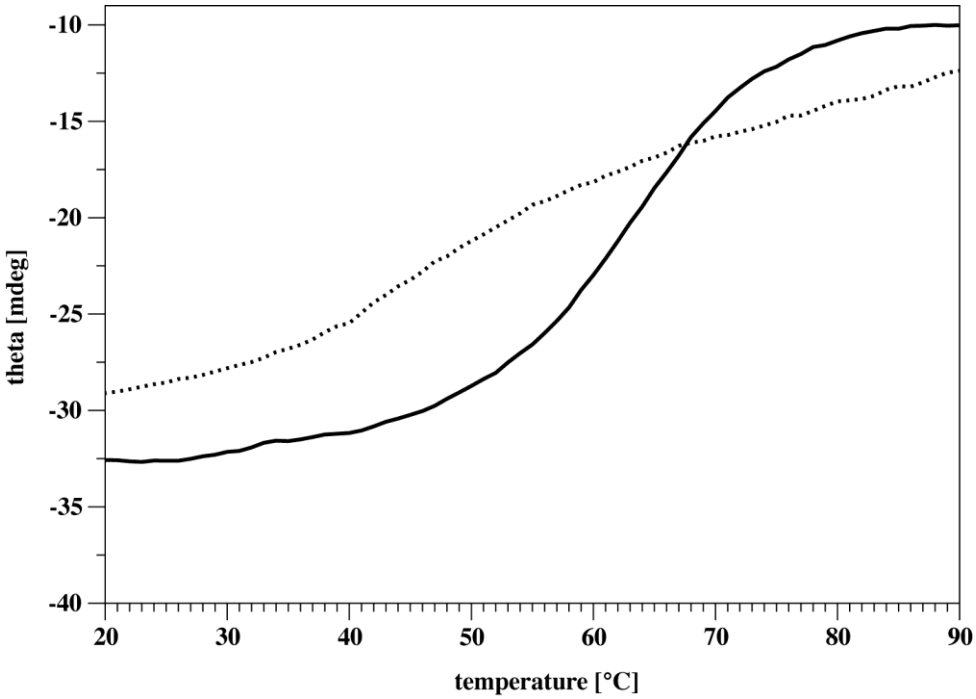


Figure S4.4 Thermal denaturation observed at 220 nm for YM_2 (dotted line) and MA (solid line)

4.6.1.7 **NMR data:**

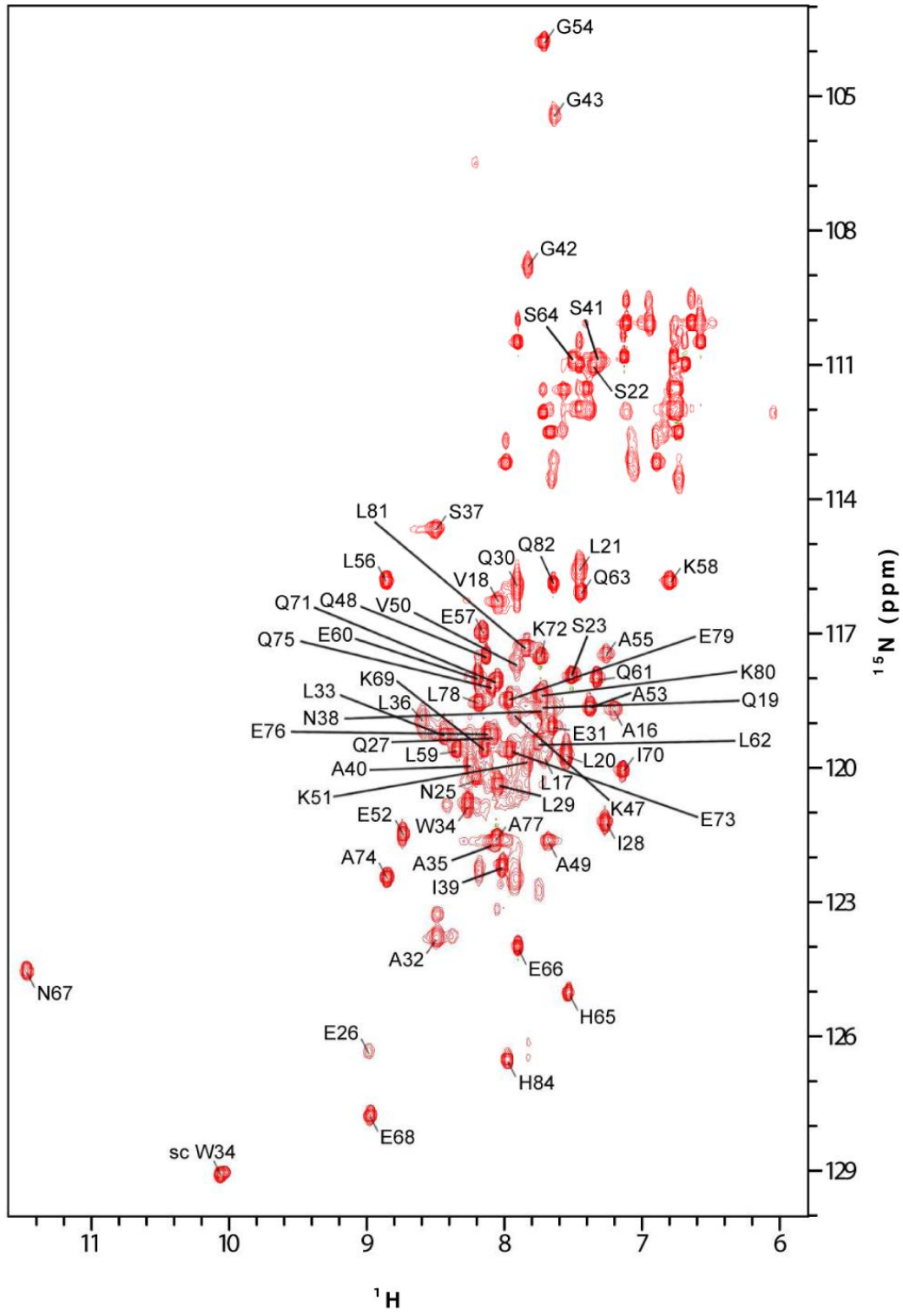


Figure S4.5 600 MHz [^{15}N,^{1}H]-HSQC spectrum of uncomplexed MA recorded at 310 K, 0.75 mM in PBS_{150} pH 7.4, 2 % glycerol, 10 % D_2O, 1 mM TMSP, 0.02 % NaN_3.

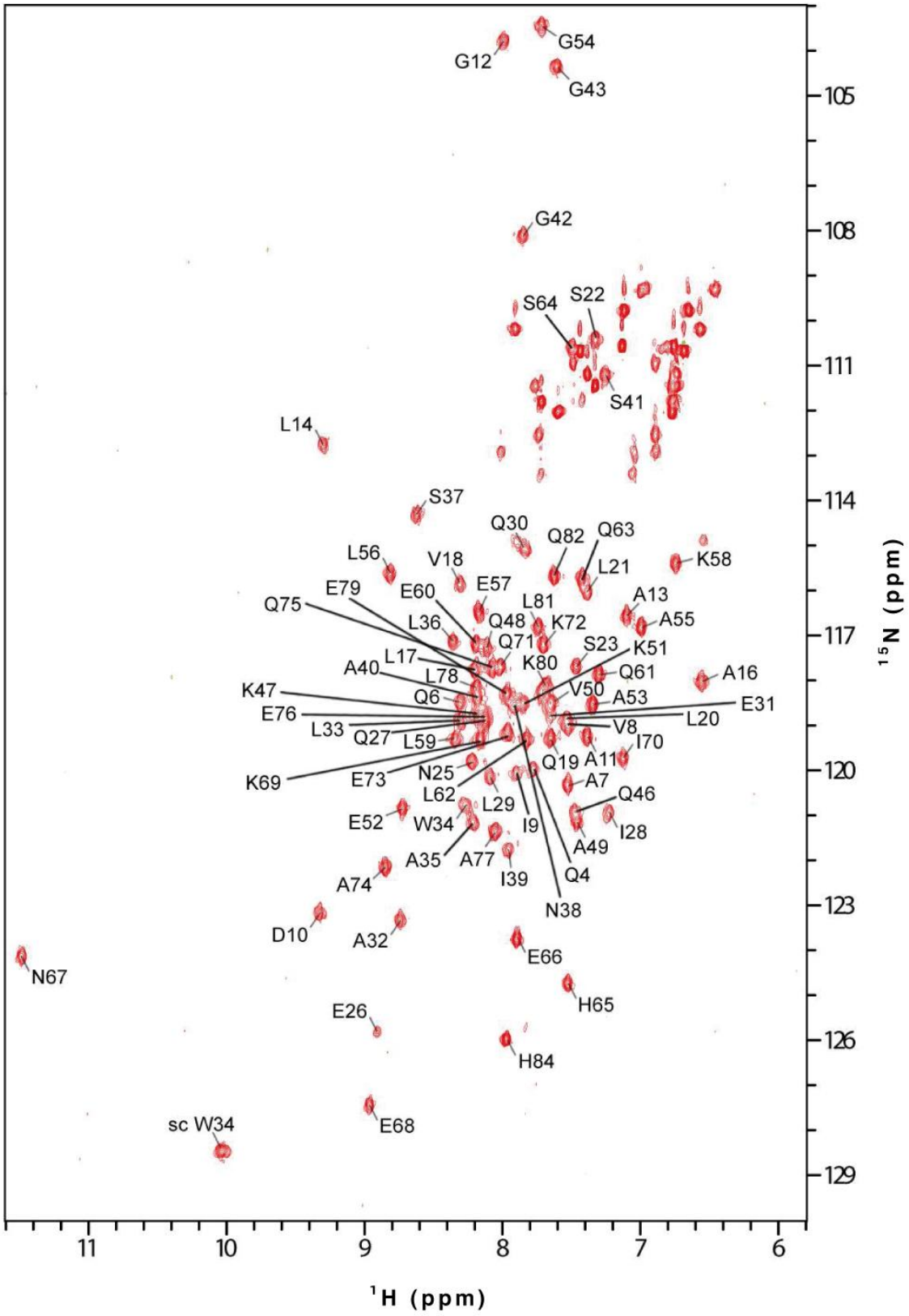


Figure S4.6 600 MHz [^{15}N,^{1}H]-HSQC spectrum of MA complexed with 1.2 equiv. of YM_2, recorded at 310 K, 0.75 mM, in PBS_{150} pH 7.4, 2 % glycerol, 10 % D_2O, 1 mM TMSP, 0.02 % NaN_3.

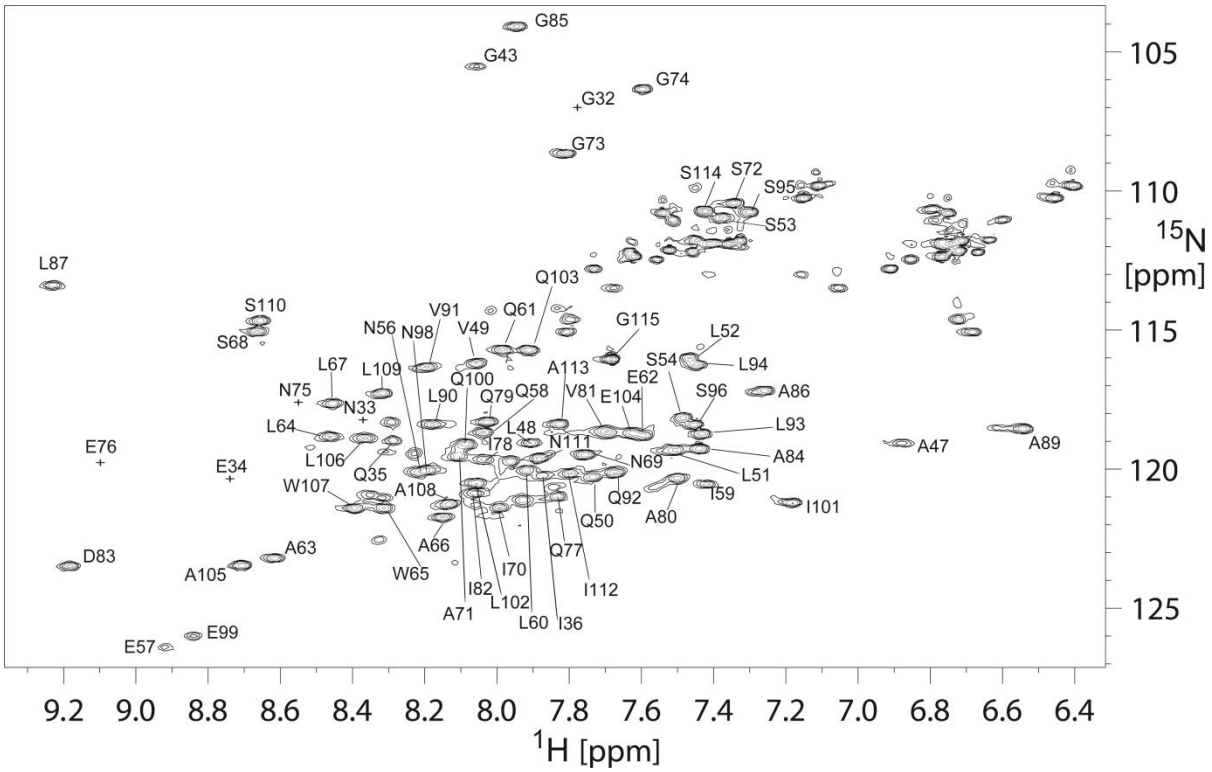


Figure S4.7 700 MHz [^{15}N,^{1}H]-HSQC spectrum of YM_2 complexed with 1.5 equiv. of MA, recorded at 310 K, 1 mM in PBS_{150} pH 7.4, 2 % glycerol, 10 % D_2O, 1 mM TMSP, 0.02 % NaN_3.

Table S4.6 Structure statistics for uncomplexed MA

Experimental Input	
Total NOE restraints	788
Intraresidual (|i-j| = 0)	242 (30.7%)
Sequential (|i-j| = 1)	237 (30.1%)
Medium-range (1 < |i-j| < 5)	167 (21.2%)
Long-range (|i-j| > 4)	142 (18.0%)
Dihedral restraints	443
RMSD of backbone heavy atoms (N, Cα and C')	
of residues 14-42 and 46-82	0.80 ± 0.21 Å
RMSD of all heavy atoms	
of residues 14-42 and 46-82	1.34 ± 0.19 Å
Target function (average)	0.99 ± 0.20 $Å^2$
Target function (lowest value)	0.68 $Å^2$
PROCHECK (Laskowski *et al.*, 1996) statistics	
Residues in most favored regions	80.3%
Residues in allowed regions	18.4%
Residues in generously allowed regions	0.0%
Residues in disallowed regions	1.3%

Table S4.7 Structure statistics for complexed MA

Experimental Input	
Total NOE restraints	715
Intraresidual (|i-j| = 0)	290 (40.6%)
Sequential (|i-j| = 1)	232 (32.5%)
Medium-range (1 < |i-j| < 5)	132 (18.5%)
Long-range (|i-j| > 4)	61 (8.5%)
Dihedral restraints	448
RMSD of backbone heavy atoms (N, Cα and C')	
of residues 6-42 and 46-84	1.09 ± 0.29 Å
RMSD of all heavy atoms	
of residues 6-42 and 46-84	1.51 ± 0.26 Å
Target function (average)	1.16 ± 0.20 $Å^2$
Target function (lowest value)	0.83 $Å^2$
PROCHECK statistics	
Residues in most favored regions	89.5%
Residues in allowed regions	9.2%
Residues in generously allowed regions	1.3%
Residues in disallowed regions	0.0%

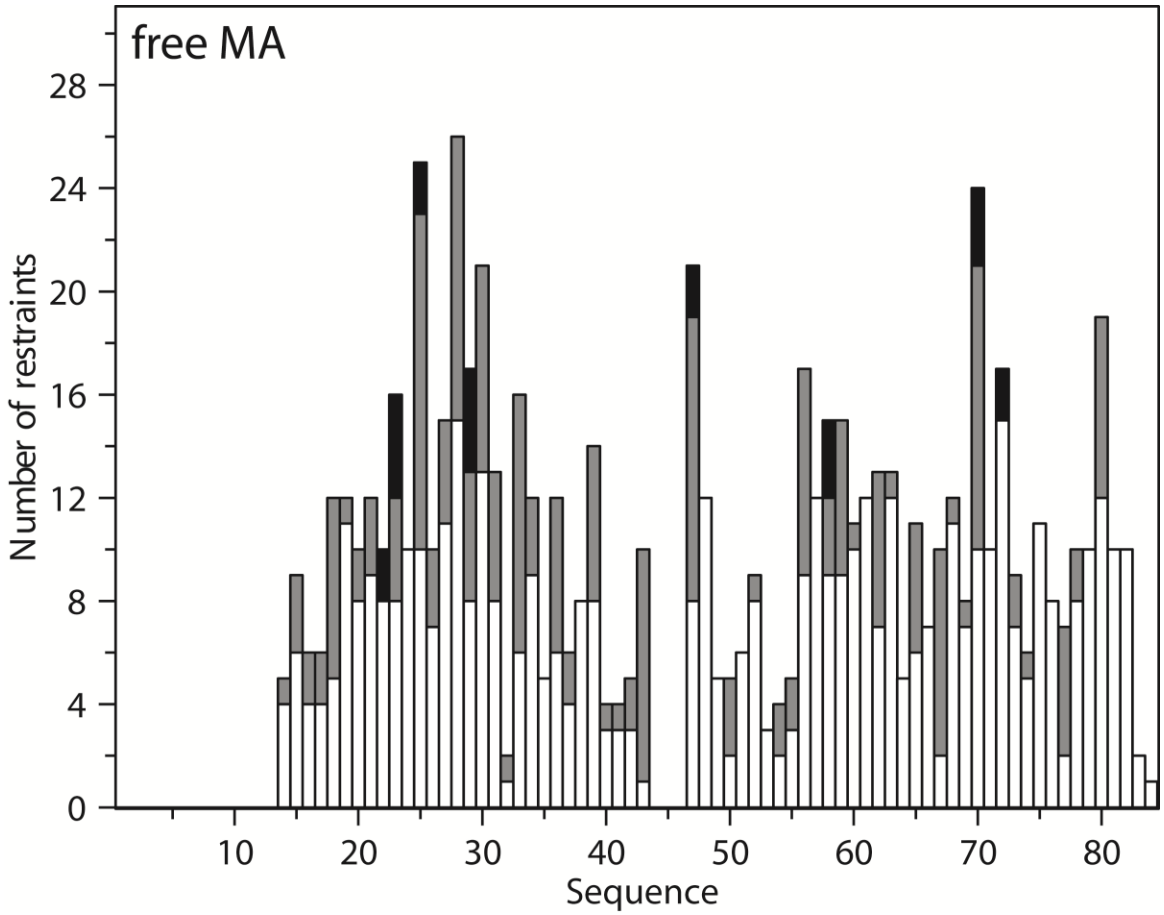


Figure S4.8 NOE restraints per residue as used in the final cycle for the calculation of the uncomplexed MA structure. Bars are further divided into intraresidual NOEs (|i-j| = 0) and sequential NOEs (|i-j| = 1) in white, medium-range NOEs (1 < |i-j| < 5) in grey and long-range NOEs (|i-j| > 4) in black.

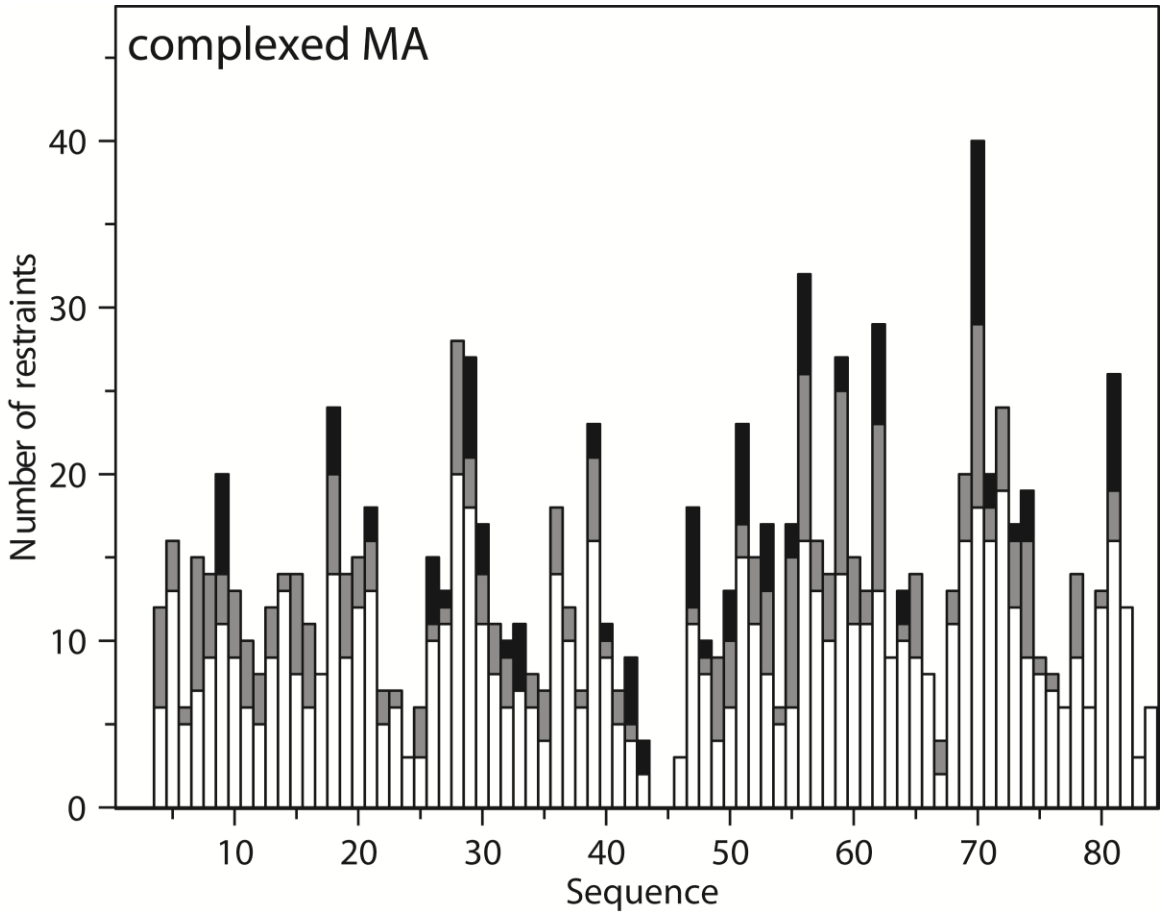


Figure S4.9 NOE restraints per residue as used in the final cycle for the calculation of the complexed MA structure. Bars are further divided into intraresidual NOEs (|i-j| = 0) and sequential NOEs (|i-j| = 1) in white, medium-range NOEs (1 < |i-j| < 5) in grey and long-range NOEs (|i-j| > 4) in black.

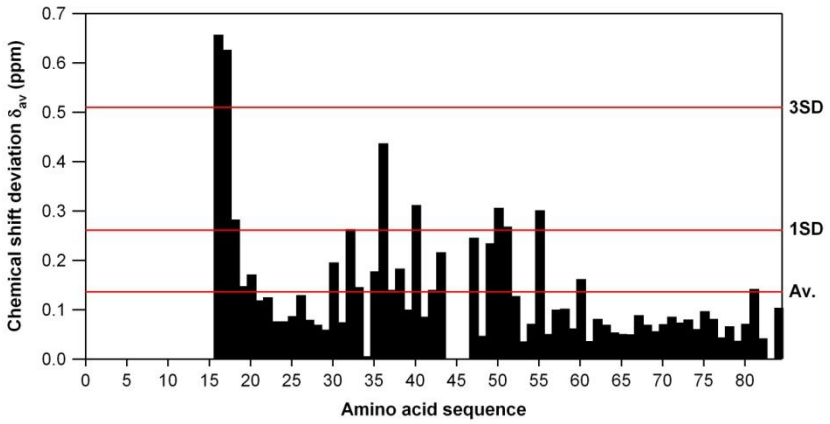


Figure S 4.10 Chemical shift deviations (δ_{av}) for MA upon complex formation with NMR-invisible YM_2. Quantification of deviations was calculated according to the formula $\delta av = \sqrt{\left((\delta HN)^2 + \left(\frac{\delta N}{5}\right)^2\right)}$ where δHN and δN correspond to the amide proton and nitrogen chemical shifts, respectively[3]. The horizontal lines mark the average deviation across all residues (0.136 ppm), plus one standard deviation (1SD) and three standard deviations (3SD).

4.6.1.8 Secondary structure prediction of YM_2 when complexed with MA

Although no high-resolution structure of the YM_2 fragment is available so far, nearly-complete backbone assignments for the complexed YM_2 except for the (invisible) N-cap are available (97 %), allowing us to predict secondary structure based on H, N, C' and Cα chemical shifts using the program TALOS+. Figure S4.11 demonstrates that the predicted secondary structure nearly perfectly coincides with the expectations from the YM_3A crystal structure (PDB # 4DBA):

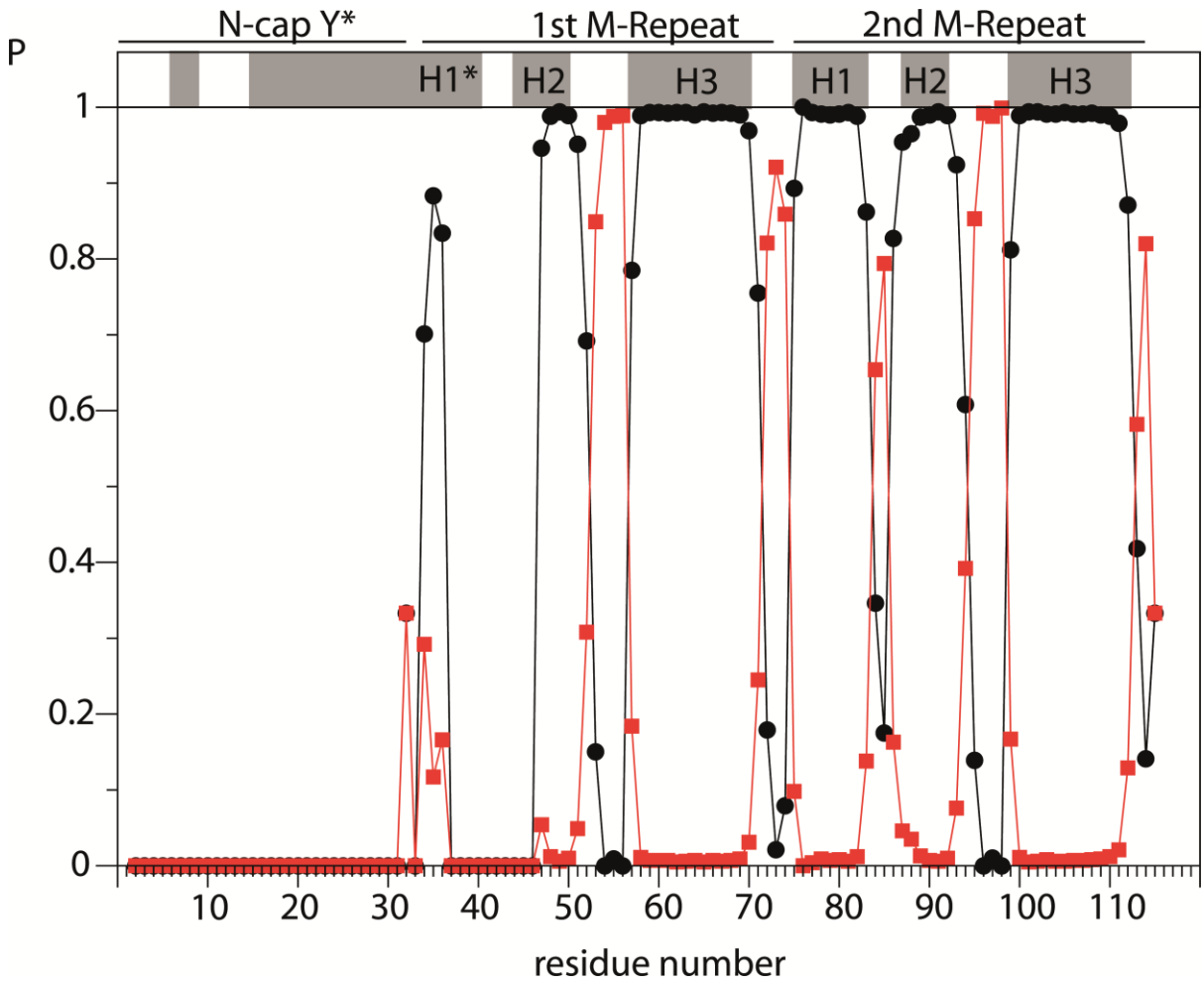


Figure S4.11 Probability (P) for formation of helical (black circles) or coiled (red squares) dihedrals using the program TALOS+. On top the location of helices are indicated by grey boxes based on the YM_3A crystal structure. Note: Residues 1-31 as well as 36-46 in the N-terminal part of YM_2 are not assigned.

***The grey boxes indicating a long extended helix containing helix one (H1) of the 1st M-repeat and parts of the N-cap (Y) are based on the crystal structure of YM_3A. A swapped dimer conformation is observed under crystallization conditions such that helix one is extended. In solution however, YM_3A is monomeric.**

In addition to the backbone assignment, the complete side-chain assignment for the second M repeat of YM_2 was achieved. This allowed a preliminary analysis of the complex interface using isotope filtered 3D NOESY experiments. To this end the software MolProb was used to generate the positions of protons on the YM_3A crystal structure[2]. Based on this a list of expected NOEs was generated assuming an initial cut off at 5 Å. Then the relevant interface positions of YM_2 were examined using ^{15}N, ^{13}C

filtered, ^{13}C edited 3D NOESY spectra of ^{15}N, ^{13}C labelled YM_2 in complex with unlabeled MA for the occurrence of the predicted NOE signals originating from the MA fragment. In a second step, relevant probable distances between protons was increased to 6 Å to allow for potentially different rotamers in the solution structure interface as compared to the crystal structure. For a preliminary set of identified interface NOEs see Table S4.8.

Table S4.8 Table of observed interface NOEs originating from MA, observed on YM_2 side.

Proton in YM_2	Proton in MA	Distance
Cut-offs		0-5 Å
78 Ile HG2	4 Glu HG2	4.77
87 Leu HD1	11 Ala HB	2.70
94 Leu HD1	20 Ala HB	4.33
102 Leu HD2	28 Ile HD1	4.03
102 Leu HD2	28 Ile HD1	2.96
106 Leu HD1	31 Glu HB3	2.54
106 Leu HD2	31 Glu HB3	1.64
106 Leu HD2	20 Ala HB	4.90
109 Leu HD1	8 Val HG2	4.54
Moiety in YM_2	Moiety in MA	Distance
Cut-offs		5-6 Å
82 Leu HD1	4 Glu HG2	5.77
91 Val HG2	17 Leu HG1	5.69
102 Leu HD2	28 Ile HB	5.08
106 Leu HD1	17 Leu HD2	5.74
106 Leu HD2	28 Ile HB	5.50
106 Leu HD2	17 Leu HD2	5.51
106 Leu HB	17 Leu HD2	5.34
109 Leu HD1	17 Leu HD2	5.87
109 Leu HD1	17 Leu HG	5.24
109 Leu HG	8 Val HG2	5.50
112 Ile HG2	8 Val HG2	5.65
113 Ala HB	5 Ile HG2	5.36

The upper part of the table shows NOEs observed within the 5 Å cut-off, the bottom part lists the additionally observed NOEs in the 5-6 Å range. Matching was obtained with a ±0.025 ppm shift difference cut-off.

* Shift difference between the frequency observed in the YM_2 data set and the frequency observed in the MA data set.

4.6.2 Supplementary References

[1] C. Aslanidis, P. J. de Jong, *Nucleic Acids Res* **1990**

[2] C. Madhurantakam, G. Varadamsetty, M. G. Grütter, A. Plückthun, P. R. Mittl, *Protein Sci.* **2012**

[3] I. Radhakrishnan, G. C. Perez-Alvarado, D. Parker, H. J. Dyson, M. R. Montminy and P. E. Wright, *J Mol Biol* **1999**).

5. An Interdisciplinary Approach to Investigate Peptide Binding to a Designed Armadillo Repeat Protein Improving Protein Design using NMR, MD and other Biophysical Techniques

Christina Ewald[1], Randall P. Watson[1], Ting Zhou[2], Martin T. Christen[1], Annemarie Honegger[2], Amedeo Caflisch[2], Andreas Plückthun[2]*, Oliver Zerbe[1]*

[1] Department of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

[2] Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

*corresponding authors:

Email address of corresponding authors: oliver.zerbe@oci.uzh.ch; plueckthun@bioc.uzh.ch

5.1. Abstract

The specific recognition of peptide sequences by proteins plays an important role both in biology and in diagnostic applications. However, production of the most commonly used biomolecular binders, antibodies, is time- and resource-consuming. We propose the alternative use of Armadillo repeat proteins (ArmRPs) that can bind extended peptide sequences in a modular and predictable fashion. Here we characterize the binding mode of neurotensin (NT) against the previously developed ArmRP VG_328 by solution NMR methods, mutational studies, molecular dynamics and other biophysical techniques. We describe assignment problems arising from the repetitive nature of the amino acid sequence, and present novel approaches to facilitate assignments. Fractional assignments obtained for VG_328 in combination with chemical shift perturbations (CSPs) allowed recognition of the repeats involved in binding. Subsequent removal of repeats not involved in making contacts to the peptide resulted in a reduced-size binder with very similar affinity for NT, for which complete backbone assignments were achieved. Paramagnetic relaxation enhancement (PRE) studies using spin-labeled NT, and binding affinities of mutants provided experimental evidence which were compared with binding poses derived from MD simulations. All NMR data were compatible with the binding mode extracted from the MD trajectory. This approach demonstrates that low-resolution NMR data were sufficient to guide further designs, and that the combination of NMR and MD data was successful in establishing the binding mode in the absence of crystallographic data for a weakly binding ligand.

5.2. Introduction

The easy access to oligonucleotides of any desired sequence via DNA synthesis has revolutionized molecular biology allowing manipulation of genetic material to become a simple routine task. Similarly, the utilization of binding proteins which bind peptides or extended parts of target proteins in a sequence-specific manner could have a comparably transforming effect on various fields such as proteomics, structural biology, medical diagnostics and even therapy.

Many proteins of interest have disordered termini or loosely packed loops, however, no binding proteins have yet been developed to allow target binding in a rational way based on a target sequence.

Currently, monoclonal or recombinant antibodies [1] and a range of other scaffolds [2–6] are available as protein or peptide binding reagents. The most prominent drawback of these systems is that for each new target a completely new binder must be established, characterized and tested for specificity. Binders for similar targets established in the past do not provide sufficient design information for future projects. Furthermore, many of these scaffolds such as designed Ankyrin repeat proteins (DARPins) preferentially bind to the surface of folded proteins. However, unfolded protein regions and peptides with specific sequences play a vital role in cellular signalling and protein trafficking and other methods of detection and manipulation which are more suited to interactions with peptide-like extended conformations must be utilized.

Antibodies bind peptides with high affinities and have been structurally well-characterized [7], but their mode of binding is not conserved. Moreover, antibodies and their derivatives contain labile disulfide bonds rendering them unsuitable for intracellular applications. In contrast to antibodies, small adaptor domains like SH2, SH3 and PDZ domains [8] usually show specific binding in a conserved fashion within one family. However, their binding affinity is weak, only short sequences are recognized, and specificity is limited. Despite their ability to recognize a wide range of targets, major histocompatibility complex proteins [9] are unattractive candidates due to their complex handling. Repeat proteins, in particular Armadillo repeat proteins (ArmRPs) [10], tetratricopeptide repeats [11], WD40 [12] proteins, HEAT repeats [13], and Ankyrin repeats [14], possess an intrinsic ability to bind peptides due to their repetitive structure resulting in well-defined surfaces that can be used for binding. ArmRPs, which are abundant

in eukaryotes [15], and are the subject of this research, often mediate protein-protein interactions and participate in a broad range of biological processes [16]. Well known examples are β-catenin, which is involved in cell adhesion and signalling [17], and importin-α, which is vital for the nucleocytoplasmic transport of proteins [18].

Repeat modules of ArmRPs typically contain about 42 amino acids [19], which are arranged into a triangle of three α-helices (H1-3). In nature 4-12 repeats are stacked beside each other forming a right-handed superhelix, the armadillo domain, which is responsible for peptide recognition [20]. The elongated hydrophobic core is protected by specialized capping modules at the N- and C-termini. Peptides are bound in extended conformation via interactions between highly conserved asparagine side chains, which lie in a groove formed by the H3 helices, and the peptide backbone [21]. Specificity is conferred by other residues of H3 interacting with side chains of the target peptide. Each repeat of the armadillo domain specifically recognizes a dipeptide subunit of the bound peptide, providing the basis for a modular approach. Dissociation constants (K_d) as low as 10-20 nM have been reported [22,23]. Designed ArmRPs based on natural consensus sequences are available from previous studies [24]. These repeat proteins have been designed for superior *E. coli* expression and application in intracellular environments, avoiding both surface exposure of large hydrophobic patches and the presence of cysteine residues in the sequence. The designed ArmRP scaffold proteins are soluble, highly expressed, stable, monomeric and display improved characteristics compared to natural ArmRPs. The original design by Parmeggiani *et al.* [25] based on a sequence consensus of the importin-α and the β-catenin families, was further improved by Alfarano *et al.* [26] using a molecular dynamics (MD) based approach. The resulting scaffold was found to be very stable, and was employed in the creation of randomized libraries by Varadamsetty *et al.* [24].

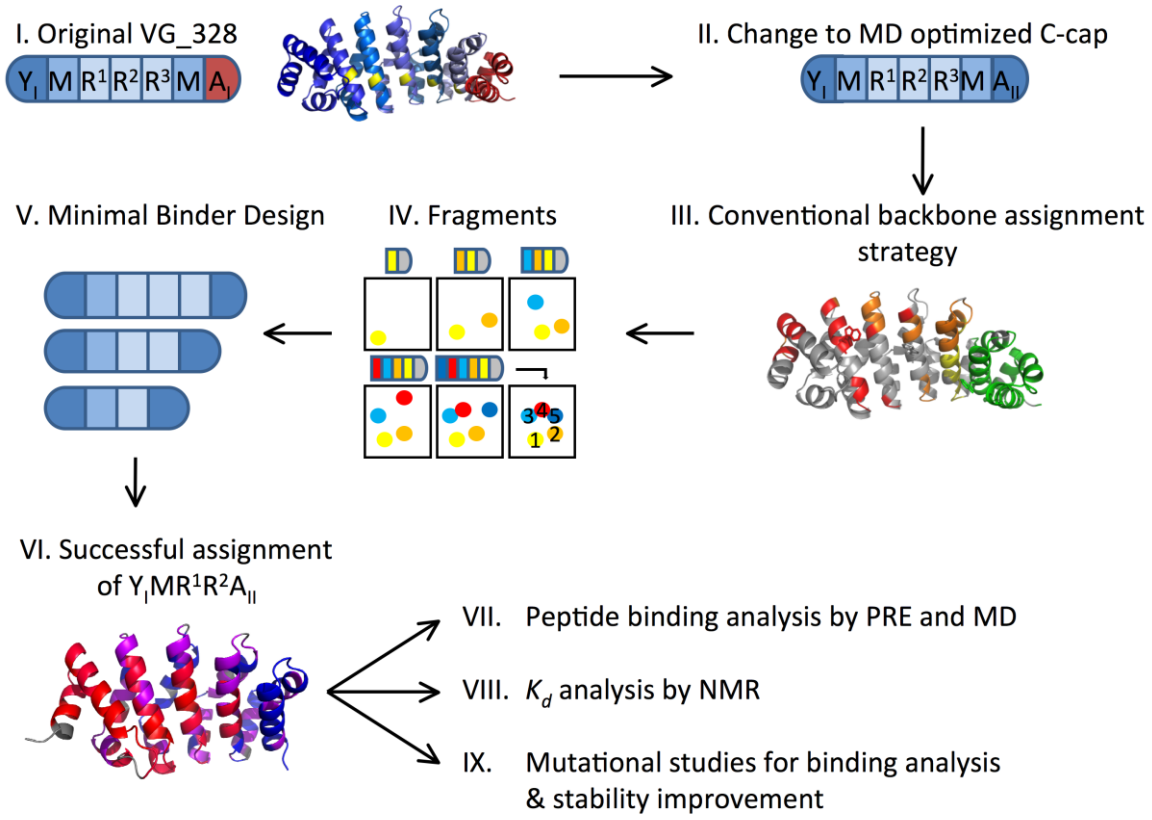


Figure 5.1 General strategy for NMR assignment and characterisation of NT peptide binding to an engineered armadillo repeat protein based on VG_328.

A member of this library, VG_328, was selected to bind the human neurotensin (NT) peptide using ribosome display. The 13-amino acid neuropeptide NT of the sequence QLYENKPRRPYIL was chosen as the target peptide for its lack of defined structure in solution [27]. The selected 32 kDa ArmRP VG_328 binds NT with a K_d of 7 µM at 4° C and contains five internal repeats flanked by N- and C-terminal capping repeats. The residues responsible for peptide binding specificity on the surface of H3 of the central three repeats have been randomized (see Figure 5.2). Using ELISA assays with single-site alanine mutants of NT, VG_328 has been shown to specifically bind NT, with four key NT side chains P7, R8, R9, and Y11 contributing to the binding (see Figure 5.9). The moderate affinity is suggested to result from only four residues of the 13-residue peptide contributing most of the binding energy [24].

Figure 5.2 The sequence of proteins based on $Y_IMR^1R^2R^3MA_{II}$ and associated modifications and mutations. Proteins were expressed in *E. coli* with one of the displayed expression sequences (red boxes) from pPANK- (expression of whole proteins and C-terminally truncated protein fragments) and pLIC-based (expression of N-terminally truncated protein fragments) expression plasmids. For example, the sequence of $Y_IMR^1R^2R^3MA_{II}$ includes the pPANK-based His_6-tag and all capping and repeat sequences indicated by its name. Shorter versions, for example $Y_IMR^1R^2A_{II}$ include the sequences of the pPANK-based His_6-tag, the Y_I N-cap, the first unrandomized M-type repeat, randomized repeats R^1 and R^2 and the A_{II} C-cap. C-terminally truncated proteins end at position 41 of their last respective internal repeat, e.g. $Y_IMR^1R^2$ ends with the glutamate in position 41 of the R^2 repeat. N-terminally truncated proteins are headed by a TEV-cleavable His_6-tag (pLIC-based) and include the number of repeats indicated by their name, e.g. R^3MA_{II}. Due to the glycine residue left after TEV-cleavage (ENLYFQ/G) at the N-terminus of the fragment, N-terminally truncated proteins effectively start with the last glycine of the preceding repeat, which is naturally present in the sequence. For cloning reasons the $MR^1R^2R^3MA_{II}$ fragment starts two residues N-terminal of the glycine adding "RD" to the sequence. Helices 1, 2 and 3 of each repeat and the C-cap are indicated by light blue boxes, the N-cap only contains two helices , H1 and H2. The three central internal repeats containing randomized positions are boundered by a dark blue line. Randomized positions in H3 of randomized internal repeats are indicated in red. Spontaneous point mutations acquired during ribosome display are indicated in green. Sites of mutations introduced in different cap generations to improve protein stability are indicated in bold and underlined: V34, R37 and R42 (continuous numbering including pPANK-based His_6-tag) are mutated to convert a Y_I N-cap into a Y_{II} N-cap

(R34, S37, R42Δ); The LQ motif at the end of the A_{II} C-cap corresponds to QF in a A_I C-cap.

In order to undertake iterative optimization of designed ArmRPs, it was pivotal to determine the binding mode of the first generation binder VG_328 to NT. The aim of the present study was to analyze this interaction in detail to guide future design efforts. Therefore, NMR studies were performed to learn more about this protein-peptide interaction, to determine the binding location and orientation of NT on VG_328, and to establish if NT binds in the canonical orientation observed in natural ArmRPs. An overview of our strategy can be found in Figure 5.1. Due to the low affinity of VG_328 for NT (approximately 7 μM) [24] the interaction was analyzed by NMR spectroscopy as it can offer detailed information, even for interactions with K_d values in the mM to μM range [28]. However, as described below in more detail, the repetitive nature of the armadillo sequence results in a number of technical challenges, some of which were already encountered in our previous studies of Ankyrin repeat proteins [29]. To facilitate this process we employed a wide range of techniques, from isotopic labeling and fragmentation of ArmRPs, to the selective deletion of repeats. The results from these studies have culminated in the design of a reduced-size binder that was much more amenable to the NMR analysis. The backbone of this minimal binder could be assigned nearly completely, and subsequent chemical shift perturbation experiments in combination with PREs from ligand-attached spin labels allowed the derivation of experimental constraints for peptide binding analysis. In combination with MD simulations a low-resolution picture of the complex emerged, that will be very useful in future design rounds.

5.3. Materials & Methods

5.3.1. Nomenclature

The ArmRPs in this study contain consensus repeats (M) and randomized internal repeats (R) based on the previously described M̄-type (for a more detailed description of the sequence nomenclature see Alfarano *et al.* 2012) [24,26]. In the case of VG_328 the protein contains 3 randomized library modules termed R^1, R^2 and R^3. The N-terminal capping repeat, derived from *y*east importin-α is termed "Y". The C-terminal capping repeat, was *a*rtificially designed [26] termed "A". The number of identical repeats in a protein is indicated as a subscript, e.g. a protein with five identical internal consensus repeats is called YM_5A. To distinguish different cap versions of capping repeats, the caps are labeled with additional subscripts in roman numerals, e.g. $Y_{II}M_5A_{II}$. In the presented nomenclature the used binder VG_328 is denoted as $Y_IMR^1R^2R^3MA_I$. All used cap and internal repeat sequences are shown in Figure 5.2.

5.3.2. Molecular Biology

Experiments were performed according to standard procedures unless stated otherwise [30]. Oligonucleotides were purchased from Microsynth AG (Balgach, Switzerland), for a complete list of all used oligonucleotides see Table 5.2. Enzymes and buffers were from New England Biolabs or Fermentas (Lithuania), and the *E. coli* strain XL1-blue (Genotype: *recA1, endA1, gyrA96, thi-1, hsdR17*(r_K^-, m_K^+), *supE44, relA1, lac,* [*F', proAB, lacIqZΔM15::Tn10(tetr)*], Stratagene, California, USA) was used for cloning. Chemically competent cells were prepared using the $CaCl_2$- method [30]. Oligonucleotides and vectors were designed in Vector NTI (Invitrogen).

Polymerase chain reactions (PCR) were usually performed with 200 µM of total dNTPs, 1.5 µM of forward and reverse primers and 50 ng of template. Taq polymerase was used for colony-PCRs, *Pfu* polymerase for site-directed mutagenesis and general high fidelity DNA amplification. Point mutations were introduced by whole plasmid amplification using non-overlapping primers, one of which introduced the mutation, and blunt end ligation.

DNA restriction, phosphorylation and ligation were performed according to the manufacturer's instructions (New England Biolabs, USA, Fermentas, Lithuania or Promega, USA). DNA quality and size was assessed on ethidium bromide stained (0.1 µg/ml) 1% agarose gels. PCR products were purified using standard DNA kits (Machery&Nagel, Germany, or Roche, Switzerland). DNA concentration was determined by UV absorbtion spectroscopy at 260 nm (ND-1000 spectrophotometer Nanodrop Technologies, USA).

pLIC_CR (Reichen *et al*. unpublished) and pPANK, a pQE30 derivative lacking the *Bpi*I and *Bsa*I endonuclease restriction recognition sites (GenBank accession number AY327140) were used as cloning and protein expression vectors. All pPANK-based vectors introduce an $MRGSH_6$-tag at the N-terminus of the proteins. pLIC-based vectors carry a TEV protease cleavable N-terminal H_6-tag (See Figure 5.2)

5.3.3. Cloning of $Y_IMR^1R^2R^3MA_{II}$ Fragments

pPANK_$Y_IMR^1R^2R^3MA_{II}$ has previously been established from pPANK_NyM328MCa [24] by introducing two point mutations (Q292L, F293Q) [26] in the A_I-type C-cap.

N-terminal fragments (Y_IMR^1, $Y_IMR^1R^2$, $Y_IMR^1R^2R^3$) were cloned by deleting the unrequired C-terminal repeats from previously established pPANK_NyM328MCa encoding $Y_IMR^1R^2R^3MA_I$. C-cap primers 053 and 056 were used to create $YMR^1R^2R^3$ (template: pPANK_NyM328MCa [24]), 054 and 056 for YMR^1 (template: pPANK_NyM328MCa [24], and 63 and 65 for YMR^1R^2 (template: pPANK_$Y_IMR^1R^2R^3$) (see Table 5.2). Primers were designed to allow amplification of the whole expression plasmid excluding the sequence to be deleted. The PCR product was then treated with *Dpn*I (Fermentas) to remove template DNA, treated with PNK T4-Kinase (Promega) and blunt end ligated to re-create a circular pPANK based plasmid.

C-terminal fragments (R^3MA_{II}, $R^2R^3MA_{II}$, $R^1R^2R^3MA_{II}$ and $MR^1R^2R^3MA_{II}$) were amplified from pPANK_$N_IM328MC_{II}$ using primers 078 and 088, (R^3MA_{II}), 077 and 088 ($R^2R^3MA_{II}$ and $R^1R^2R^3MA_{II}$) and primers 090 and 088 ($MR^1R^2R^3MA_{II}$). Primers were designed to allow ligation independent cloning (LIC) [31] into the pLIC_CR vector (Reichen et al. unpublished) to introduce a cleavable N-terminal His_6-tag for IMAC purification and undisturbed fragment interaction. pLIC_CR contains an N-terminal $MRGSH_6$-tag followed by the TEV protease recognition site and the fatal *sacB* gene as an additional selection marker. pLIC_CR_MA for the expression the C-terminal MA has been previously established by Watson *et al*. (Watson *et al*. unpublished [32]). For LIC cloning 100 ng of pLIC_CR was *Bsa*I digested, and treated with T4-DNA-polymerase in the presence of dCTP. The target PCR fragment was treated with T4-DNA-polymerase and dGTP. Prepared plasmid and insert DNA were column purified (High Pure PCR Product Purification kit, Roche), mixed at approximately 1:1 molar ratio and incubated for 10 min at room temperature and then transformed into *E. coli* XL1-blue cells. Plasmids obtained from colonies on LB agar plates containing 100 µg/ml ampicillin and 7 % sucrose were sequenced for correctness.

5.3.4. Cloning of $Y_IMR^1R^2R^3A_{II,}$ $Y_IMR^1R^2A_{II}$ and $Y_IMR^1A_{II}$

Vectors for $Y_IMR^1R^2R^3A_{II,}$ $Y_IMR^1R^2A_{II}$ and $Y_IMR^1A_{II}$ were cloned by deleting the unrequired internal repeats from $Y_IMR^1R^2R^3MA_{II}$ via the same blunt end ligation strategy described above for the cloning of the N-terminal fragments. pPANK_$Y_IMR^1R^2R^3A_{II,}$ was amplified from pPANK_$Y_IMR^1R^2R^3A_{II}$ using primers 101/054. pPANK_$Y_IMR^1R^2A_{II}$ and pPANK_$Y_IMR^1A_{II}$ were amplified from pPANK_$Y_IMR^1R^2R^3A_{II}$ using primers 101/063 and 101/056 respectively.

5.3.5. Cloning of $Y_IMR^1R^2A_{II}$-Mutants

The following single, double and triple point mutations in pPANK_$Y_IMR^1R^2A_{II}$ were introduced by whole plasmid amplification and blunt end ligation as described above using primers indicated in Table 5.2: V34R, R37S, R42Δ, R42A, E46A, V34R/R37S, V34R/R42Δ, R37S/R42Δ and V34R/R37S/R42Δ (= $Y_{Ii}MR^1R^2A_{II}$).

5.3.6. Expression of Unlabeled and Isotopically Labeled Proteins

Unlabeled ArmRPs and ArmRP fragments were expressed in *E. coli* M15 [pREP4] cells (K12 derivative, genotype: *NalS strS rifS thi*$^-$ *lac*$^-$ *ara*$^-$ *gal*$^+$ *mtl*$^-$ *F*$^-$ *recA*$^+$ *uvr*$^+$ *lon*$^+$ [*pREP4 KanR*], Qiagen, Germany) containing the pREP4 plasmid, which co-expresses the *lac* repressor. Cells were transformed with the target plasmid and grown at 37 °C in LB medium containing 50 µg/L kanamycin and 100 µg/L ampicillin in baffled 2 L Erlenmeyer flasks at 37 °C shaking at 220 rpm. When cultures reached OD_{600} = 0.6 expression was induced with IPTG at 1 mM final concentration. Cells were harvested by centrifugation at 9000 g for 20 min after 5 h. Cell pellets were stored at -20 °C.

^{15}N ^{13}C-labeled proteins were expressed starting from 5 mL LB overnight cultures, as described above, which were centrifuged at 2000 g for 10 min to remove unlabeled media. Cells were carefully re-suspended in minimal media containing $^{15}NH_4Cl$ and ^{13}C-glucose (Sigma-Aldrich, Switzerland, see Table Table 5.3) and supplemented with 1 x trace metal solution (see

Table 5.4). Antibiotic concentrations in minimal media were reduced to 50 µg/L ampicillin and 12.5 µg/L kanamycin and the induction time was increased to 16 h.

Expression of ^{15}N, ^{13}C, ^{2}H-labeled proteins was performed in the same manner as described above for ^{15}N, ^{13}C-labeling. Antibiotic concentrations in deuterated minimal media were further reduced to 15 µg/L ampicillin and 12.5 µg/L kanamycin. However, to adapt the culture to growth in deuterated medium, the 5 mL overnight starter cultures contained LB

prepared with D_2O instead of H_2O. This was used to inoculate an additional 50 mL D_2O minimal medium pre-culture, which was incubated overnight to increase cell density before being used to inoculate the final culture at a volumetric ratio of 1:20. Expression was induced at OD_{600} = 0.6 and carried out for 16 h at 37 °C. Using ^{1}H, ^{13}C-glucose the achieved level of deuteration was about 90 %.

5.3.7. Protein Purification and Characterization

Cell pellets were thawed, resuspended in TBS_{500} [50 mM Tris-HCl, 500 mM NaCl, 5 % (v/v) glycerol, pH 8.0] and lysed by sonication on ice (Branson digital sonifier, Model 250 with microtip, Missouri, USA; 20 % amplitude, pulse length 10 s, rest period 15 s for 10 min total sonication). Insoluble cell debris were pelleted by centrifugation for 30 min at 17,000 g. The filtered supernatant (0.2 µm cellulose acetate syringe filters, Sarstedt) was purified by Immobilized Metal Affinity Chromatography (IMAC) using Hi-Trap Ni columns (Pharmacia) pre-equilibrated with TBS_{500}. After washing the column with 10 column volumes (CV) of TBS_{500} and 3 CV of TBS_{500} containing 5 mM imidazole to remove unspecifically bound contaminants, the target ArmRP protein was eluted with TBS_{500} containing an adequate concentration of imidazole (25-200mM; target dependent).

To remove the His_6-tag of C-terminal ArmRP fragments for complexation, TEV(SH) protease [33] was added to IMAC eluates at a molar ratio of 1:30 and dialyzed against 200 x volume of PBS_{150} (50 mM Na-phosphate, 150 mM NaCl, 2 % (v/v) glycerol, pH 7.4) at room temperature. The cleavage was monitored by SDS-PAGE, and when the reaction was complete, the solution was filtered (0.2 µm cellulose acetate syringe filters, Sarstedt) to remove precipitated TEV(SH). Cleaved His_6-tag and His_6-tagged TEV(SH) were removed by a further IMAC step (Hi-Trap Ni column, Pharmacia), the eluate now containing only the cleaved target.

After IMAC purification ArmRPs and fragments were further purified by preparative size exclusion chromatography (SEC) in PBS_{150} pH 7.4 with 2 % (v/v) glycerol on a S75 16/60 HiLoad (GE Healthcare) at 1 mL/min before being used for further experiments. After preparative SEC protein solutions were concentrated using spin concentrators (Centricon filters, Millipore) for further analysis. Protein size and purity were checked by 15% SDS-PAGE, stained with Coomassie Blue R (GE Healthcare, Switzerland). Proteins were further analyzed by ESI mass spectrometry to verify the exact mass and determine the

degree of isotopic labeling. Extinction coefficients were calculated using the Protparam tool of the ExPASy proteomics server to determine protein concentrations by absorbance at 280 nm (ND-1000 spectrophotometer Nanodrop Technologies, USA).

Analytical SEC was carried out on a Superdex 200 5/150 GL (Pharmacia) column on an ÄKTA HPLC system. Injections of 50 µL were run in PBS_{150} (50 mM phosphate and 150 mM NaCl, pH 7.4, 2 % glycerol) at a flow rate of 0.3 mL/min. Calibration was carried out with a mixture of 2 MDa Blue Dextran, 67 kDa albumin (17-0442A, Pharmacia LMW std), 44.3 kDa ovalbumin (A-5378, Sigma, Chicken egg), 25 kDa chymotrypsinogen (17-04542B, Pharmacia LMW std) and 13.7 kDa ribonuclease A (R-5503, Sigma, from bovine pancreas). ArmRPs have been shown to elute at a higher apparent size than the calculated monomeric weight suggests. This is a result of their elongated shape and greater effective hydrodynamic ratio [24,34].

5.3.8. NMR Spectroscopy and Data Evaluation

All proteins were analyzed in PBS_{150} buffer (150 mM NaCl, 50 mM Na-phosphate, 2 % (v/v) glycerol, pH 7.4) supplemented with 10 % D_2O, 1 mM TMSP-d4 and 0.0 1% NaN_3. The presence of 2 % (v/v) glycerol in the buffer significantly improved NMR sample stability over extended periods at 310 K (data not shown). Protein solutions were concentrated to 0.2-1.0 mM for NMR measurements (Centricon filters, Millipore). NMR data were recorded at 310 K on Bruker AV-600 or AV-700 MHz spectrometers equipped with triple-resonance cryoprobes. TOPSPIN 2.1 was used for data processing using mirror-image linear prediction for constant-time evolution periods. CARA (Keller 2004) was used for further evaluation and resonance assignment. Resonances were calibrated relative to the proton water resonance at 4.63 ppm, the ^{15}N and ^{13}C scales were calculated indirectly (conversion factors $^{15}N = 0.10132900$, $^{13}C = 0.25144954$). Experiments were selected from the Bruker standard pulse sequence library, and used pulsed-field gradients, sensitivity-enhancement schemes, and water suppression through coherence selection. Proton-nitrogen and proton-carbon correlation maps were derived from [^{15}N,^{1}H]-HSQC and [^{13}C,^{1}H]-HSQC experiments.

5.3.9. Backbone and Side Chain Assignment

For backbone assignments, ^{15}N, ^{13}C, ^{2}H-labeled proteins were used. Deuterium decoupling was applied during relevant ^{15}N- or ^{13}C-evolution periods or delays. Sequential amide spin

systems were linked via matching carbonyl (HNCO/HN(CA)CO experiments) and Cα and Cβ resonances (HNCACB/HN(CO)CACB experiments). Additionally, HN(CACO)NH and ^{15}N-3D-NOESY experiments provided sequential correlations of nitrogens and protons of amide groups, respectively [29]. For side chain assignments [^{13}C,^{1}H]-HSQC experiments combined with (H)CCH-TOCSY and ^{13}C-resolved aliphatic or aromatic-NOESY experiments of uniformly ^{15}N, ^{13}C-labeled protein were used.

Predictions for backbone torsion angle estimates were calculated with TALOS+ [35] based on the experimentally obtained and assigned Cα and Cβ chemical shifts. These were used to confirm the correct localization of helices and loops in structural models developed *in silico*.

5.3.10. Chemical Shift Mapping (CSM) Experiments

Chemical shift mapping was used to probe for conformational changes in the protein upon peptide binding and to investigate direct protein-peptide interactions. Shift deviations (Δ_{av}) for $Y_IMR_1R_2R_3MA_{II}$ and $Y_IMR_1R_2A_{II}$ upon complex formation with 2 equivalents of NT were taken from [^{15}N, ^{1}H]-HSQC spectra and quantified using the formula $\Delta v_{obs} = \sqrt{\left((\Delta v_{HN})^2 + \left(\Delta v_N \cdot \left|\frac{\gamma_N}{\gamma_H}\right|\right)^2\right)}$, where ΔHN and ΔN correspond to the amide proton and nitrogen chemical shift differences respectively, and chemical shifts are weighted according to their gyromagnetic ratios γ_H 42.576 and γ_N.-4.3156

5.3.11. Determination of Dissociation Constants (K_d) by [^{15}N,^{1}H]-HSQC based CSM Titrations

Ligand binding was detected from perturbations of [^{15}N,^{1}H]-HSQC spectra by monitoring the chemical shift changes of the backbone amide as a function of ligand concentration. To determine the K_d from CSPs a total of 5 equivalents of 10 mM NT or NT7-13 peptide solution were successively added to 250 µM protein samples in PBS_{150} buffer (150 mM NaCl, 50 mM Na-phosphate, 2 % (v/v) glycerol, pH 7.4) supplemented with 10 % D_2O, 1 mM TMSP-d_4 and 0.0 1% NaN_3. [^{15}N,^{1}H]-HSQC spectra of 7 titration steps (0, 0.25, 0.5, 1, 1.5, 3 and 5 equivalents, taking dilution effects into account) were recorded on a Bruker AV-600 spectrometer equipped with a triple-resonance cryoprobe at 310 K. Chemical shift changes for both previously assigned resonances and unassigned resonances in both the ^{1}H and ^{15}N dimensions were monitored over the course of the titration. From the quadratically weighted amplitudes of ^{1}H and ^{15}N chemical shift differences combined chemical shift changes were calculated in ppm for each step *i*. as

$$\Delta v_{obs}^{i} = \sqrt{\left(\Delta v_{H}^{i}\right)^{2} + \left(\Delta v_{N}^{i} \cdot \left|\frac{\gamma_N}{\gamma_H}\right|\right)^{2}}$$

and plotted as a function of peptide concentration. Data were fitted by non-linear regression analysis to the theoretical curve for single-site binding described by[36]

$$\Delta v_{cal}^{i} = \Delta v_{\infty} \frac{\left([\mathrm{P}]_{\mathrm{total}} + [L]^{i} + K_D\right) - \sqrt{\left([\mathrm{P}]_{\mathrm{total}} + [\mathrm{L}]^{\mathrm{i}} + K_D\right)^2 - 4[\mathrm{P}]_{\mathrm{total}} \cdot [L]^{i}}}{2[\mathrm{P}]_{\mathrm{total}}}$$

using a custom designed MatLab script [37] in order to determine the K_d. For a system in fast exchange the theoretical $\Delta\delta_{cal}$ is determined by the total protein concentration $[P]_{total}$, $[L]^i$ denotes the current ligand concentration at each titration step.

Experimental chemical shift uncertainties, δ_{noise} were determined in NMRpipe [38]. The difference between experimentally measured $\Delta\delta_{obs}$ and calculated $\Delta\delta_{cal}$ for each titration curve was minimized as described by Christen *et al.* [37] and the uncertainties of the optimized parameters were estimated via Monte Carlo resampling as described by Webb *et al.*[39].

Multiple binding curves derived from different peaks were determined for each protein and averaged to yield a reliable K_d value. Ligand saturation between 70 % (K_d > 200 µM) and >95 % (K_d < 30µM) were achieved in the last titration step.

5.3.12. Determination of Dissociation Constants (K_d) by Surface Plasmon Resonance (SPR)

SPR was carried out on a BIACORE 3000 instrument (GE Healthcare Biosciences, Pennsylvania, USA) with PBS-T [50-mM phosphate, 150-mM NaCl, 0.01 % Tween-20, pH 7.4] as running buffer. 10 RU (response units) of synthetic, biotinylated NT were immobilized on a streptavidin-coated SA-chip (GE Healthcare Biosciences). Interactions of NT with $Y_IMR_1R_2R_3MA_{II}$ and $Y_IMR_1R_2A_{II}$ were measured at increasing concentrations of protein (0.06-200 µM, flow rate 50 µl / min, 50 µl injections, 5 min dissociation buffer flow). Measured values were corrected by subtraction of a reference signal from an uncoated cell. Due to fast equilibration of the system, plateau values were used to determine the dissociation constant (Scrubber, BioLogic software).

5.3.13. Paramagnetic Relaxation Enhancement (PRE) Experiments

The cysteine mutants of NT (NT_Q1C, NT_K6C and NT_L13C, Anaspec, see Figure 5.9, bottom) were dissolved in PBS_{150} (50 mM Na-phosphate, 150 mM NaCl, pH 7.4) and incubated with a 2 x excess of TCEP for 30 min at room temperature. A 10 x excess of the PRE-tag MTSL (CAS: 81213-52-7,TRC, Toronto) dissolved in DMSO was added, and the pH adjusted to 9 using 1 M NaOH. The reaction mix was incubated for 2 h in the dark at room temperature with vigorous shaking. Complete labeling was confirmed by ESI mass spectrometry. Labeled peptides were purified by SEC using a 30/10 peptide column (GE Healthcare) run in deionised water. Relevant fractions were lyophilized and checked by mass spectrometry for purity and size. The resulting peptides NT_Q1C-MTSL, NT_K6C-MTSL and NT_L13C-MTSL were dissolved in PBS_{150} (50 mM Na-phosphate, 150 mM NaCl, pH 7.4) and added at 2 x molar excess to NMR samples containing uniformly ^{15}N, ^{13}C-labeled $Y_IMR^1R^2A_{II}$. Two sets of [^{15}N,^{1}H]-HSQC (water flip-back) and [^{13}C,^{1}H]-HSQC (aliphatic and aromatic) experiments were recorded using relaxation delays of 2 s. The second experiment served to provide the reference values, and was recorded after addition of 10 equivalents of ascorbic acid to quench the paramagnetic nitroxide moiety of MTSL. After the addition of ascorbic acid the sample was incubated at room temperature for 1 h. Before recording the reference spectrum the pH was readjusted to pH 7.4 using 1 M NaOH. Previously assigned signals in proton-nitrogen and proton-carbon correlation maps were integrated in CARA. The ratio of the signal intensity $MTSL_{active} : MTSL_{inactive}$ was used as an indicator of spatial proximity of the PRE-tagged peptide side chain to the attenuated residues of the protein.

Similarly, the amine reactive PRE-tag Oxyl-1-NHS (CAS: 37558-29-5, TRC, Toronto) was used to label NT at its native lysine side chain in position 6 creating NT_K6-NHS. Due to the absence of cysteines in natural NT the reduction step with TCEP was omitted. The main difference between NT_K6-NHS and NT_K6C-MTSL is the different length of side chains tethering the PRE-tag to the peptide. The NMR-active PRE moiety is, in both cases, a stabilized nitroxide radical.

5.3.14. ELISA Assays

In the ELISA assays MaxiSorp 96-well plates (Nunc) were coated with NeutrAvidin (100 µl per well, 66 nM, overnight, 4 °C). The wells were blocked with 300 µl of 1×PBS-TB (50 mM phosphate and 150 mM NaCl, pH 7.4, 0.3 % BSA, 0.1 % Tween-20) for 1 h at room

temperature. The biotinylated target peptide ([Biotin]-[6-amino-caproic acid]-[β-Ala]$_2$-NT) was immobilized (100 µl per well, 200 nM, 1 h, 4 °C) in PBS-TB. Proteins were dissolved in PBS-B (50 mM phosphate and 150 mM NaCl, pH 7.4, 0.3 % BSA), and all washing steps were carried out in PBS-TB. Plates were incubated with target protein (100 µl per well, 200 nM, 1 h, 4 °C). Wells were washed three times with 300 µl of 1× PBS-TB and incubated with anti-RGSH$_6$ mouse antibody (1:5000 in 1× PBS-TB, 1 h, 4 °C; Qiagen, Germany) as primary antibody. Plates were washed as described above and incubated with a goat anti-mouse IgG alkaline phosphatase conjugate (1:10,000 in 1× PBS-BT, 1 h at 4 °C, Sigma) as secondary antibody. Signals were developed with the substrate disodium 4-nitrophenyl phosphate (100 µl per well, 3 mM, 2 h, 37 °C, Fluka, in 50 mM $NaHCO_3$ and 50 mM $MgCl_2$). Absorbance at 405 nm was measured with a Perkin Elmer HTS 7000 Plus plate reader (Reference absorbance wavelength 540 nm was deducted).

5.3.15. Molecular Dynamics Simulations

MD simulations were carried out in explicit water, at constant temperature (330 K) and constant pressure (1 atm) using a v-rescale thermostat and Berendsen pressure coupling [40,41]. The long-range electrostatic interactions were treated by the particle mesh Ewald method with 10 Å cut-off [42]. Van der Waals interactions were truncated at a 9 Å cut-off, and a switch function was activated starting at 8 Å. The LINCS algorithm was used to fix the length of all bonds [43]. Virtual sites were used for removing fastest degrees of freedom, which allowed an integration time step of 5 fs. All simulations were performed in the Gromacs program [44], with the OPLS force field [45] and the TIP3P potential for water molecules [46]). The protonation state of the side chains was chosen to reproduce the experimental pH 7.4: aspartate and glutamate side chains and the C-terminal carboxyl group were negatively charged, lysine and arginine side chains and the N-terminal amino group were positively charged, and histidine residues were kept neutral.

All structural models shown in this work were established based on PDB coordinates of experimental crystal structures of natural and designed ArmRPs by sequence adaptation, repeat merging and relaxation in Rosetta [47]. The model for $Y_IMR^1R^2R^3MA_{II}$ is based on the natural yeast karyopherin-α structure (PDB ID: 1EE4 [48]), the model for $Y_IMR^1R^2A_{II}$ is based on the designed consensus ArmRP $Y_{III}M_3A_{II}$ (PDB ID: 4DB6 [49]).

5.4. Results

5.4.1. Overview

Parmeggiani *et al.* designed a consensus sequence based on the ArmRP β-catenin and importin-α subfamilies, which yielded soluble, monomeric, highly-expressed and stable designed ArmRPs with improved characteristics compared to their natural predecessors [25]. Based on the scaffold designed by Parmeggiani *et al*. the mutant ArmRP VG_328 was selected from a randomized library against the human neurotensin (NT) using ribosome display. VG_328 was found to bind the 13-residue peptide NT with low affinity but high specificity. Four key peptide residues were identified to contribute most of the binding energy after an alanine-scan of the peptide [24], namely proline 7, arginines 8 and 9, and tyrosine 11. The study presented here aimed to investigate this protein-peptide interaction in more detail in order to characterize the exact binding mode.

None of the NT binders based on VG_328 yielded crystals of sufficient quality for structural determination by X-ray crystallography. We therefore decided to establish the binding mode using solution NMR methods. This task, however, required the development of new tools for assigning proteins with highly repetitive amino acid sequence. The development of these tools resulted in new insights of how the protein could be reduced in size while maintaining binding affinity. Chemical shift perturbations (CSPs) and paramagnetic relaxation enhancements (PREs) indicated which parts of the protein form contacts with NT. It also indicated that NT most likely does not associate in a single unique mode. Finally, a set of mutagenesis experiments in combination with NMR and MD techniques allowed the identification of contributions from residues in the N-cap.

The combined use of biophysical and biochemical techniques in a tour-de-force helped to improve properties of the binder and to gain insight into its folding properties. In what follows we describe how we improved the properties of the original binder VG_328 to make it amenable to detailed NMR studies. We then report on the attempts to solve the assignment problem inherent in repeat proteins using N-terminally truncated versions. Knowledge from these protein truncations and from CSP data derived from VG_328 were then employed in turn to design reduced-size binders. Finally, we present mutagenesis data to deconvolute

contributions from individual N-cap residues to NT binding and use MD calculations to probe NT binding and the behavior of the N-cap.

5.4.2. Stabilization of VG_328 for NMR Studies

All NMR methods that are suitable to establish the binding mode of a peptide require at least backbone and possibly also side chain chemical shift assignments of the binder. This is usually performed using $^{15}N,^{13}C$-labeled proteins and triple-resonance experiments. In addition, for efficient assignment of resonances of proteins of that size (32 kDa) perdeuteration is mandatory. The original library selected VG_328 ($Y_IMR^1R^2R^3MA_I$) contained the N- and C-terminal capping sequences developed by Parmeggiani *et al.* [25], and the ^{15}N-labeled species displayed a good-quality [$^{15}N,^1H$]-HSQC spectrum. However, the perdeuterated protein tended to form oligomers, and quickly precipitated rendering it unsuitable for solution NMR studies. Follow-up work by Alfarano *et al.* indicated that consensus ArmRPs could be significantly stabilized by introducing two mutations (originally described as Q240L, F241Q [26]) in the C-cap to form the A_{II}-cap.

We found that transferring these mutations (Q292L and F293Q) to VG_328 creating $Y_IMR^1R^2R^3MA_{II}$ did not significantly affect peptide binding in [$^{15}N,^1H$]-HSQC titration studies *(vide infra)*. We also found that $Y_IMR^1R^2R^3MA_{II}$ was sufficiently stable in perdeuterated form for NMR studies. While the C-cap mutant retained binding of NT, N-cap mutations suggested to further improve stability (mutations characterized by N_{II}-cap) [26] also abolished any binding to the protein in NMR titrations of $Y_{II}MR^1R^2R^3MA_{II}$ with NT (data not shown). We therefore continued our spectroscopic studies using the C-terminally stabilized species $Y_IMR^1R^2R^3MA_{II,}$ (see Figure 5.2 for sequence).

Two sets of triple-resonance experiments, using samples with and without the peptide, were recorded for assignment purposes. The spectra were generally of very good quality (see Figure 5.20), however, complete backbone assignments proved very challenging, partially due to substantial peak overlap in the central region of the [^{15}N, ^1H]-HSQC and, more importantly, due to degeneracy of ^{13}C chemical shifts between residues at identical positions in the different repeats. During assignment we followed a strategy described by Wetzel *et al.* [29], which uses a combination of HNCACB/HN(CO)CACB, HNCO/HN(CA)CO spectra supported by correlations from the ^{15}N-resolved NOESY. Moreover, many additional useful correlations were observed in the HN(CACO)NH experiment that often allowed gaps in the assignment to be closed.

Due to the repetitive nature of the internal repeats not many signals outside of the C-cap and helix 3 of the R^3 repeat could be unambiguously assigned. In the absence of peptide, signals from residues of the N-cap were completely missing, which indicated that residues in this part of the molecule are in conformational exchange. Interestingly, some of these signals were observed in the spectra of the complex with NT. It is unclear whether this effect is due to a direct conformational stabilization by peptide binding, or whether it stems from changes in the neighbouring internal repeat that allows better packing of the N-cap against the first internal repeat. Randomized positions provided valuable assignment anchors and allowed unambiguous assignment of some protein segments of the putative peptide-binding surface. However, in the non-randomized helices 1 and 2 many assignment fragments were found that could not be unambiguously mapped onto the sequence. (see blue fragments in Figure 5.6).

In the absence of NT the extent of the final backbone assignment for $Y_IMR^1R^2R^3MA_{II}$ was 20.6 % (for all non-proline residues and excluding the flexible histidine tag), and comprised 92.7 % of the C-cap and 94.7 % of helix 3 of the R^3-repeat.

In the presence of 2 equivalents of NT, signal dispersion improved significantly allowing the assignment of residues in the N-cap. A total assignment of 25.4 % was achieved, with 92.7 % of the C-cap, 38.7 % of the N-cap and 100% of H3 of R^3assigned respectively. Although all spectra were of very high quality it soon became clear that no more assignments were possible using this construct when uniformly labeled.

5.4.3. Truncation of $Y_IMR^1R^2R^3MA_{II}$ Aids in Backbone Assignment and Reveals Contributions of Individual Repeats to Protein Stability

As described in the last section, chemical shift degeneracy due to the repetitive sequence hampered assignments. We rationalized that if it were possible to truncate the protein by one repeat at a time this might enable us to follow signals, ultimately allowing deconvolution of the spectra into contributions from the individual repeats. Accordingly, a set of N- and C-terminally truncated fragments was designed by splitting $Y_IMR^1R^2R^3MA_{II}$ between position 41 and G42 of individual internal repeats (see Figure 5.2). The entire series of N-terminally truncated proteins contained five fragments ($MR^1R^2R^3MA_{II}$, $R^1R^2R^3MA_{II}$, $R^2R^3MA_{II}$, R^3MA_{II} and MA_{II}).

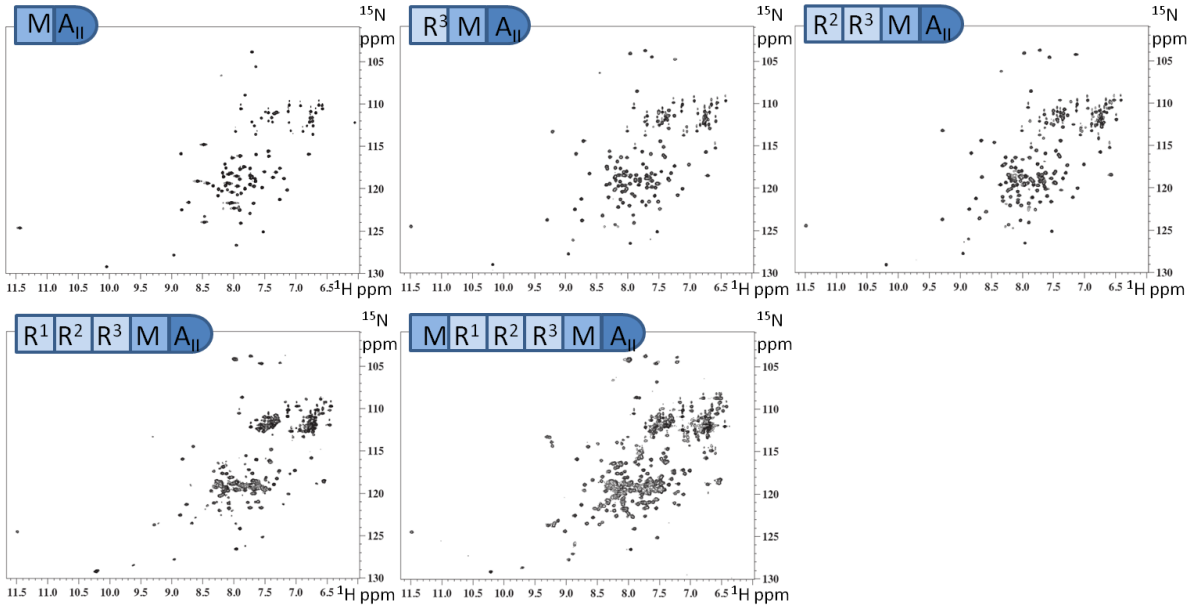


Figure 5.3 [^{15}N,^{1}H]-HSQC spectra of 300 µM N-terminally truncated $Y_IMR^1R^2R^3MA_{II}$ fragments. $R^1R^2R^3MA_{II}$ displays lower signal to noise ratio and broader line widths.

All fragments were expressed solubly and purified well. Representative [^{15}N, ^{1}H]-HSQC spectra are depicted in Figure 5.3 indicating that these fragments constitute well-folded proteins. Only spectra of $R^1R^2R^3MA_{II}$ displayed line-broadening indicating the formation of oligomeric species. The presence of three consecutive randomized repeats seems to be unfavourable; the addition of one unrandomised consensus repeat in the $MR^1R^2R^3MA_{II}$ fragment results in significantly better spectra. Parallel studies revealed that longer C-terminally truncated versions ($Y_IMR^1R^2R^3M$, $Y_IMR^1R^2R^3$, $Y_IMR^1R^2$ and Y_IMR^1) were well expressed but generally unstable and not well-folded (data not shown).

[^{15}N,^{1}H]-HSQC spectra of the ^{15}N-labeled N-terminally truncated fragments showed considerable overlap with the spectrum of full-length $Y_IMR^1R^2R^3MA_{II}$. We hypothesized that signals from each repeat would successively appear close to, or directly at, their final position over the course of extending the length of the fragment. We were easily able to transfer the available assignments of the C-cap and the last M-repeat obtained for $Y_IMR^1R^2R^3MA_{II}$ to the MA_{II} fragment. Signals were then tracked from the shortest to the longest variant. Signals were assigned to repeats by the order of their appearance in spectra of fragments of increasing size. Figure 5.4 depicts the assignment strategy and shows an example of signals successively appearing in the glycine region of the spectra.

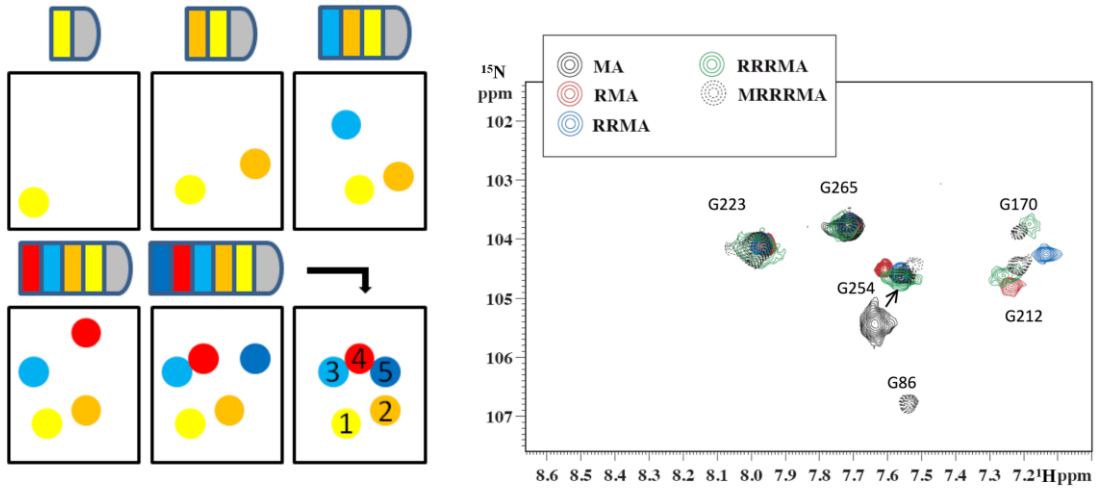


Figure 5.4 The assignment strategy using successive addition of repeat modules utilising the spectra of the N-terminally truncated fragments (left). Glycine residues in an overlay of [^{15}N,^{1}H]-HSQC spectra of 300 µM N-terminally truncated $Y_IMR^1R^2R^3MA_{II}$ fragments showing the successive appearance of signals from different repeats. For example, the signal of G254 which is located at the N-terminus of the MA_{II} fragment initially displays a different chemical shift, but moves towards a relatively stable range of chemical shifts as soon as adjacent repeats are added in longer fragments. On the other hand the signal of G86 only appears in the spectrum of the longest fragment $MR^1R^2R^3MA_{II}$.

This strategy extended our assignment significantly, and stretches of sequence previously identified as part of an unrandomized helix 2 could now be mapped to a specific repeat. Most importantly we were able to distinguish signals from the two identical helices 3 of the two unrandomized M repeats, and assigned at least parts of helix 3 in all repeats. We observed that signals of residues close to the truncation site tended to move into their final position only after another repeat module had been added in contrast to residues further away from the truncation site. We were able to efficiently employ this strategy up to $R^2R^3MA_{II}$, after which the increased spectra complexity and broad lines of $R^1R^2R^3MA_{II}$ prevented reliable transfer of assignments. We therefore back-tracked assignments in the opposite direction from the full-length $Y_IMR^1R^2R^3MA_{II}$ to $MR^1R^2R^3MA_{II}$ in order to close the assignment gaps as far as possible. Using this method we were even able to assign the side chain indole protons of the three Trp residues present on the binding surface in repeats 1, 3 and 5 without recording

specific spectra for side chain assignments. For $Y_IMR^1R^2R^3MA_{II}$ without NT we achieved a backbone assignment coverage of 36.8 % improving from the original 20.6 %. For $Y_IMR^1R^2R^3MA_{II}$ in complex with NT 44.9 % (originally 25.4 %) was obtained. These initial assignments were used to identify protein regions directly or indirectly affected by the binding of NT. [^{15}N, ^{1}H]-HSQC-based chemical shift mapping experiments of $Y_IMR^1R^2R^3MA_{II}$ with NT (see Figure 5.5) revealed that, in agreement with its moderate K_d, the system is in fast exchange. Signal broadening at intermediate titration steps was observed.

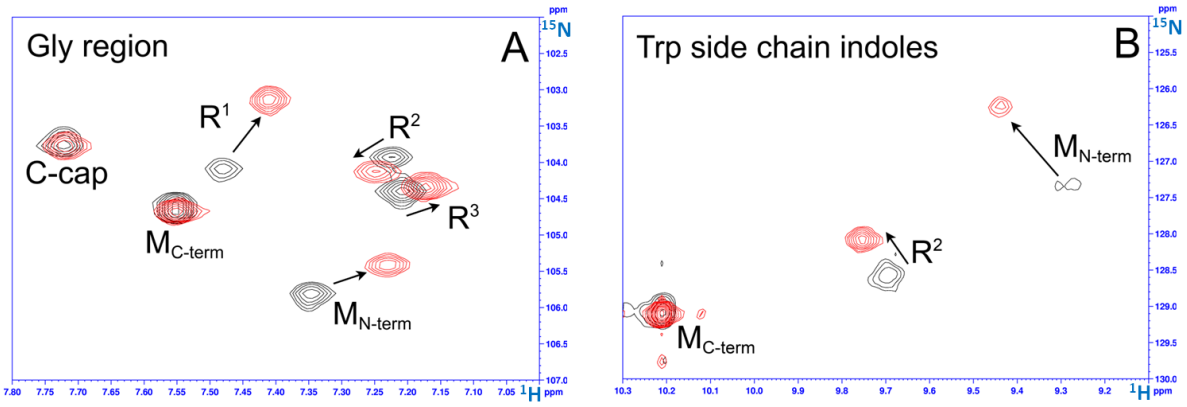


Figure 5.5 Chemical shift perturbations in [^{15}N,^{1}H]-HSQC spectra of 400 µM ^{15}N-labeled $Y_IMR^1R^2R^3MA_{II}$ without (black) and with two equivalents NT (red). The Gly (A) and indole Trp (B) regions show changes in the different repeats in detail. For the complete spectrum see Figure 5.20.

A gradient in magnitude of CSP was observed across the protein. Larger changes occur in the N-terminal region and across helices 3 of the first three internal repeats whereas the C-terminus and the fifth internal repeat are not affected. Similarly, the CSPs for the Trp indole protons are large in repeats 1 and 3, but the Trp in repeat 5 is unaffected. Figure 5.6 depicts CSPs of $Y_IMR^1R^2R^3MA_{II}$ upon addition of two equivalents of NT, mapped onto a structural model. The model was computed based on natural yeast karyopherin-α (PDB ID: 1EE4 [48]) and was energy-minimized using Rosetta [47].

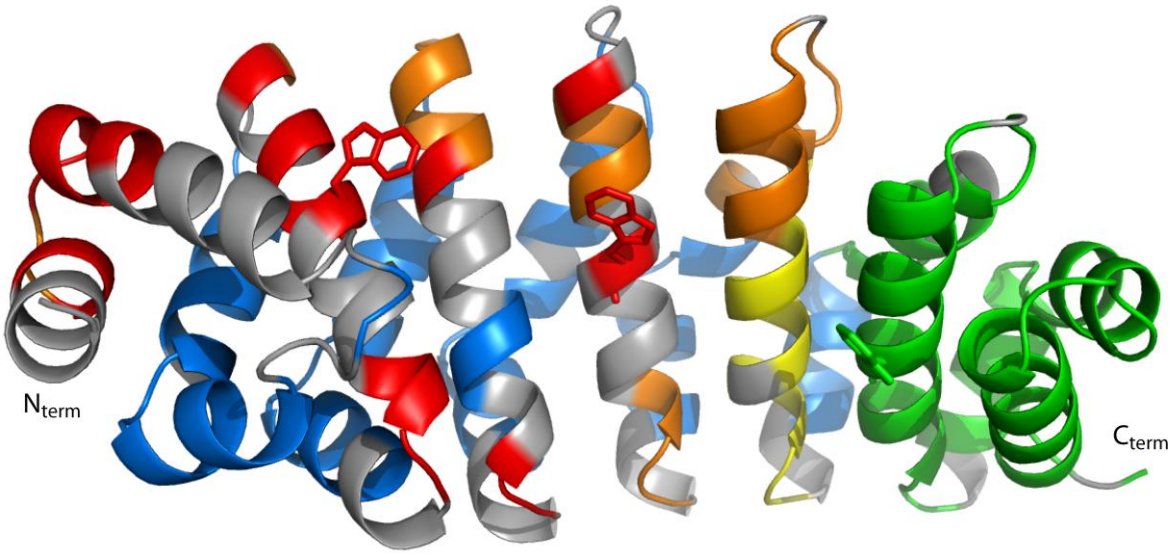


Figure 5.6 Chemical shift perturbations of $Y_IMR^1R^2R^3MA_{II}$ upon addition of two equivalents of NT mapped onto a structural model, based on natural yeast karyoperin-α (PDB ID 1EE4). Color coding: unassigned residues (grey), large shift changes (> 1 line width (lw) or untraceable: red), limited shift changes (0.5-1 lw, orange), minor shift changes (0.2 -0.4 lw, yellow), no shift change (green). Ambiguous assignments: blue. Note: For better visibility tryptophan side chains have been coloured completely to indicate CSPs of indoles.

5.4.4. Design of the Minimal-Size Binder $Y_IMR^1R^2A_{II}$ Reduces Target Complexity

The interaction studies with the truncated proteins quickly revealed that not all repeats of the original binder were involved in forming contacts to the peptide. Unfortunately, these truncated versions were not sufficiently stable, largely because they missed the N-terminal cap, which shields the hydrophobic core against solvent. In order to assess binding of NT in the context of more stable constructs than the truncated versions, $Y_IMR^1R^2R^3MA_{II}$ was successively minimized by removing internal repeats starting from the C-terminal consensus repeat (M) while retaining the stabilizing C-capping repeat (see V. in Figure 5.1 and sequences in Figure 5.2). The established ArmRPs $Y_IMR^1R^2R^3A_{II}$, $Y_IMR^1R^2A_{II}$ and $Y_IMR^1A_{II}$ were all soluble, well-expressed, and yielded high quality [^{15}N, ^{1}H]-HSQC spectra (see Figure 5.7). CSP studies of these proteins revealed that even the shortest construct $Y_IMR^1A_{II,}$ containing only two internal repeats, displayed some, albeit strongly attenuated, binding affinity (see Figure 5.7). $Y_IMR^1R^2A_{II}$ was the shortest (22 kDa) construct displaying an affinity for NT that is comparable to $Y_IMR^1R^2R^3MA_{II}$ (32 kDa). In addition, the chemical

shift perturbation pattern and degree (for the complete spectrum see Figure 5.21) were almost identical, and hence further investigations were carried out with this minimal binder.

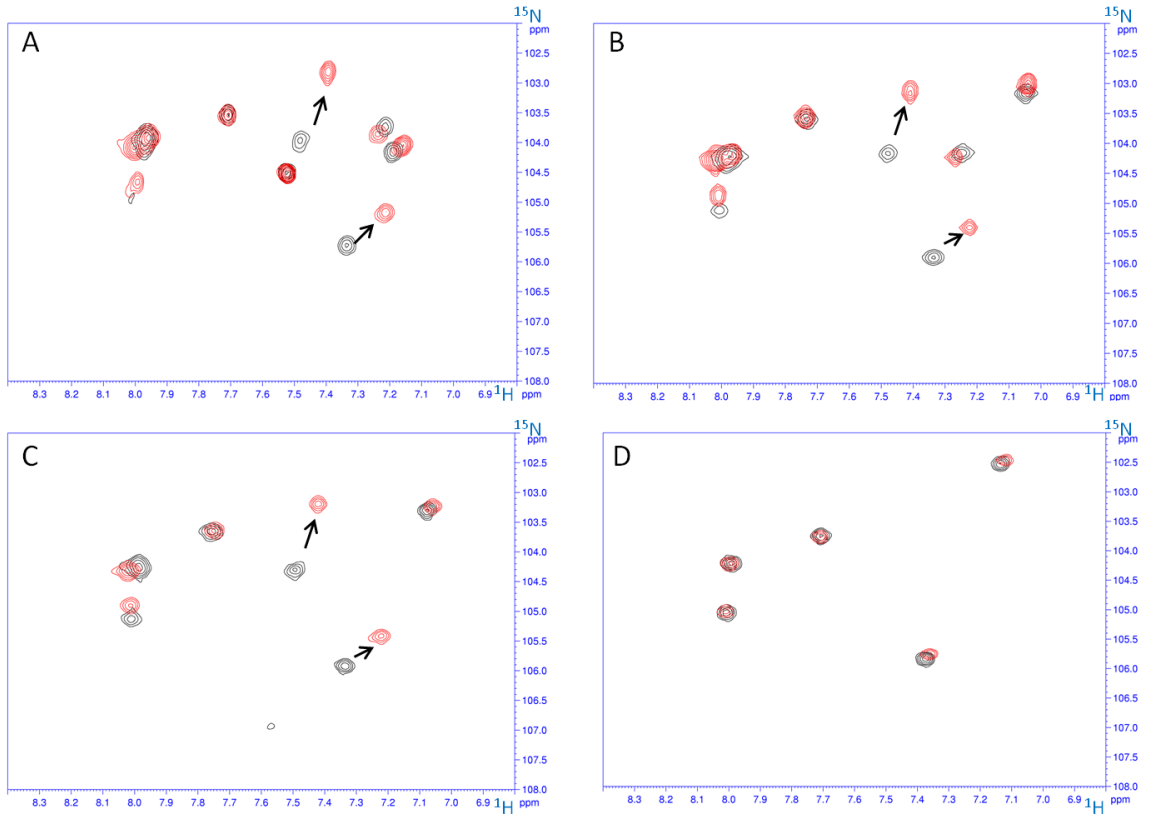


Figure 5.7 Expansion of the Gly region in [^{15}N,^{1}H]-HSQC spectra of 400 µM ^{15}N-labeled $Y_IMR^1R^2R^3MA_{II}$ (A), $Y_IMR^1R^2R^3A_{II}$ (B), $Y_IMR^1R^2A_{II}$ (C), $Y_IMR^1MA_{II}$ (D) without (black) and in complex with two equivalents NT (red). $Y_IMR^1R^2A_{II}$ is the smallest protein still showing native-like chemical shift changes (C).

A full set of triple-resonance experiments similar to the one recorded for full-length $Y_IMR^1R^2R^3MA_{II}$ was measured in both the absence and presence of NT. Assignments in this protein were achieved with much less ambiguity, largely due to the reduced size and the absence of a second identical M repeat resulting in better spectra with significantly less peak overlap. In this case almost all backbone resonances of H3 could be assigned. In these experiments, some resonances from the N-cap were observed in the free form for the first time, and in the NT-complex nearly the complete N-cap could be assigned. For $Y_IMR^1R^2A_{II}$ 82.8 % assignment coverage of backbone resonances was achieved. In complex with NT the coverage increased to 97.4 % so that only five residues in loop regions and at the beginning of the N-cap remained unassigned (see grey residues Figure 5.8).

To facilitate locating the NT-binding interface within $Y_IMR^1R^2A_{II}$ side chains of the putative binding interface were partially or fully assigned. To this end a set of amide-anchored triple-resonance spectra of $Y_IMR^1R^2A_{II}$ in complex with NT was recorded such as (H)CC(CO)NH and H(CCCO)NH [50]. These spectra allowed us to directly link side chain resonances to the already assigned backbone moieties. Additionally, $^{13}C,^1H$-based spectra were used to confirm the obtained assignments such as ^{13}C-resolved NOESY centered on aliphatic and aromatic resonances as well as (H)CCH-TOCSY spectra. The HB(CBCG)CDHD spectrum was used to link already assigned Cβ resonances of aromatic side chains to the Cδ. As a result, nearly complete assignment of the $[^{15}N,^1H]$-HSQC spectra and partial assignment of the $[^{13}C,^1H]$-HSQC spectra were now available for peptide interaction studies using CSPs and PRE probes.

5.4.5. Interaction Studies of $Y_IMR^1R^2A_{II}$ and NT Using CSP and PRE Data, and MD Simulations

Nearly complete backbone assignments in both the free protein as well as in the complex with NT allowed the evaluation of the CSP data. Figure 5.8 depicts the wide range of chemical shift perturbations observed in the N-cap and across the expected binding interface formed by helices 3 of all repeats. The observation that residues 16-18, 22, 23, 26, and 32-42 of the N-cap can only be detected by NMR in the presence of NT indicate that the N-cap is in intermediate conformational exchange in the absence of NT, and the newly visible residues become locked into one conformation upon binding of the peptide. Therefore, the CSPs result from both direct protein-peptide interactions and indirect effects due to ligand-triggered conformational changes.

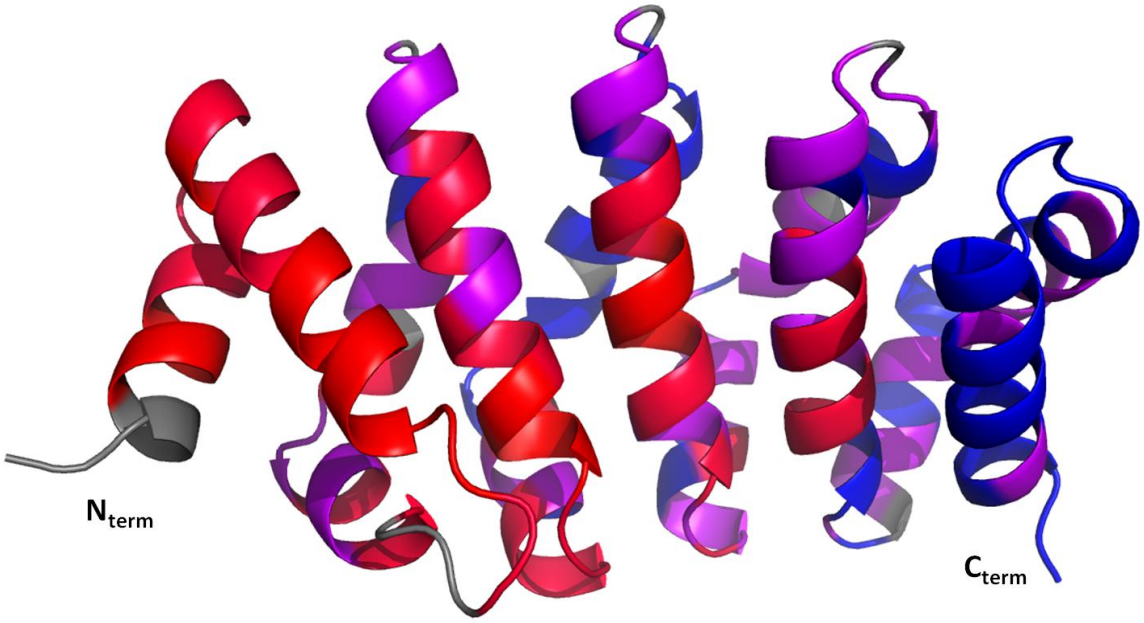


Figure 5.8 CSPs for $Y_IMR^1R^2A_{II}$ upon NT binding mapped onto a structural model based on the unrandomized designed armadillo repeat protein $Y_{III}M_3A_{II}$ (PDB ID 4DB6). Colour coding: unassigned residues or Pro (grey), large shift changes (> 1 line width (lw) or untraceable, red), limited shift changes (≤ 1 lw, violet), no shift change (blue).

To deconvolute direct and indirect effects in the CSP experiment and to more precisely locate the NT binding site we initially tried to probe more directly for peptide-protein interactions in $Y_IMR^1R^2$ R^3MA_{II} using ^{15}N,^{13}C-isotope edited/filtered NOESY experiments. Unfortunately we were unable to detect intermolecular NOEs, most likely due to the moderate affinity preventing efficient buildup of NOEs between protein and peptide in the bound state. This view is supported by the observation that the complex is in fast exchange on the NMR timescale. We therefore chose to determine the binding location and orientation of NT on $Y_IMR^1R^2A_{II}$ using paramagnetic relaxation enhancement (PRE) tags attached to NT. In this way, the magnitude of attenuation of previously assigned protein signals is related to the proximity of the PRE-tagged peptide. This effect, unlike the chemical shift changes, is entirely related to distance proximity. NT was labeled with nitroxyl PRE-tags in three different positions at the N- and C-terminus as well as at a central position (Positions 1, 6 and 13, Figure 5.9, bottom), avoiding interference with any of the four peptide residues previously identified as critical for binding (positions 7, 8 ,9 and 11) [24]. For comparison we also created a model of the complex based on the crystal structure of $Y_{III}MMMA_{III}$ [49] through homology

modelling and energy minimization using the program Rosetta [47]. Attenuation from the N-terminal spin label was very weak in comparison to the other spin label positions. Interestingly, the attenuation pattern of the PRE-tag positions 1 and 13 were similar and the effect diffused over the binding interface, whereas labeling of NT at position 6 lead to distinct and strong attenuation in the upper part of the binding interface (see Figure 5.9). The obtained results confirmed that NT binds to the binding interface formed by the third helix H3 of each repeat. They also indicated that NT may not, at least temporarily, be fully extended when bound to $Y_I MR^1R^2A_{II}$. We also noticed a patch of attenuated residues in the vicinity of residue G44 in the loop connecting the N-cap to the first repeat. This patch is distant to the putative binding interface, and cannot be explained by the structural model presented in Figure 5.9.

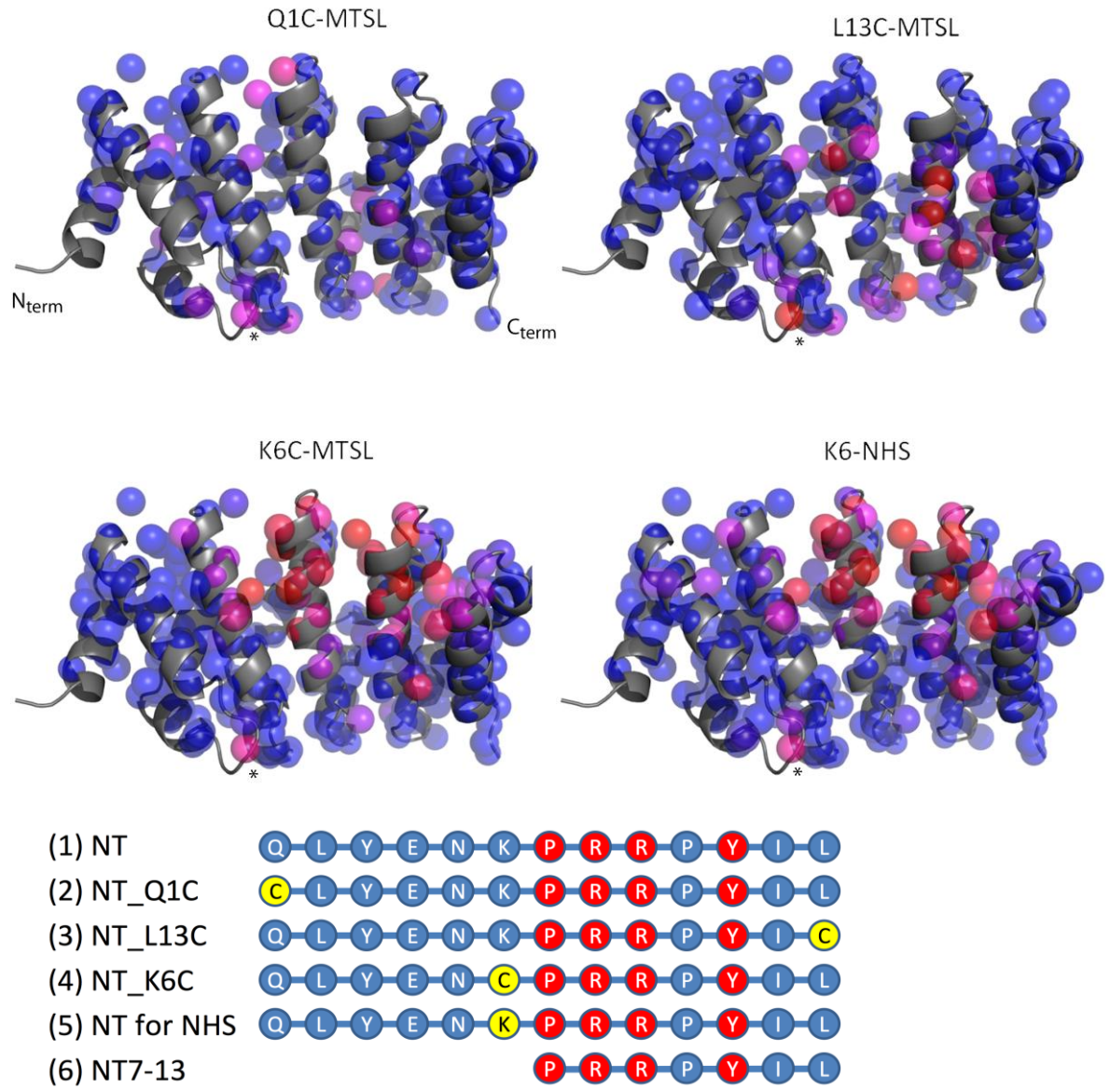


Figure 5.9 Top: Signal attenuation observed in PRE experiments of $Y_IMR^1R^2A_{II}$ with spin labelled NT peptides mapped onto a structural model based on the unrandomized designed armadillo repeat protein $Y_{III}M_3A_{II}$ (PDB ID 4DB6). Colour code: A colour gradient from red to blue depicts signal attenuation from 100 % (signal intensity fully attenuated) to 0 % (unaffected). G44 is marked with *. Bottom: List of peptides used for CSP (1 and 6) and PRE studies (2-5). The spin labels MTSL (2-4) or NHS (5) were coupled to the cysteine and lysine peptide residues indicated by yellow circles. Peptide termini were uncapped except for the acetylated N-terminus of NT7-13 (6). Note: in solution the free N-terminal glutamine of the peptide is fully converted into the cyclic pyroglutamate form. For detailed peptide sequences see

Table 5.5. Red circles indicate peptide residues identified to be of importance for NT binding by alanine-scanning [24].

Protein variants with sequences of internal repeats identical to $Y_IMR^1R^2R^3MA_{II}$ but containing stabilized versions of the N-cap are not capable of binding the NT peptide (*vide supra*). NMR data indicate that the N-cap is not well folded, and the absence of signals is indicative of molten-globule type behavior. Taking into account that previous MD simulations of proteins with this N-cap have indicated that it does not pack well against the remainder of the protein [26] we decided to calculate 1 μs MD trajectories of $Y_IMR^1R^2A_{II}$ with explicit water. The simulation was carried out at 330 K (for details see Materials and Methods) with the above-described structural model as the starting conformation. The results of the MD calculations are summarized in Figure 5.10. We observed that the N-cap and the loop containing residues 38-48, which connect the N-cap and the first internal repeat, had considerable flexibility including detachment of the N-cap from the internal repeats. Additionally a rotation of the cap against the internal repeats along the axis of the elongated hydrophobic core of the protein was observed. Through the rotation of the entire N-cap the loop connecting the N-cap with the first repeat is transferred into closer proximity to the binding surface possibly accounting for the PRE attenuations around residue G44. The same type of calculation carried out for the unrandomized consensus protein $Y_{III}M_3A_{III}$ with stabilized capping repeats revealed that the Y_I-cap exhibits significantly more flexibility than the more stable Y_{III} version of the N-cap (see Figure 5.10).

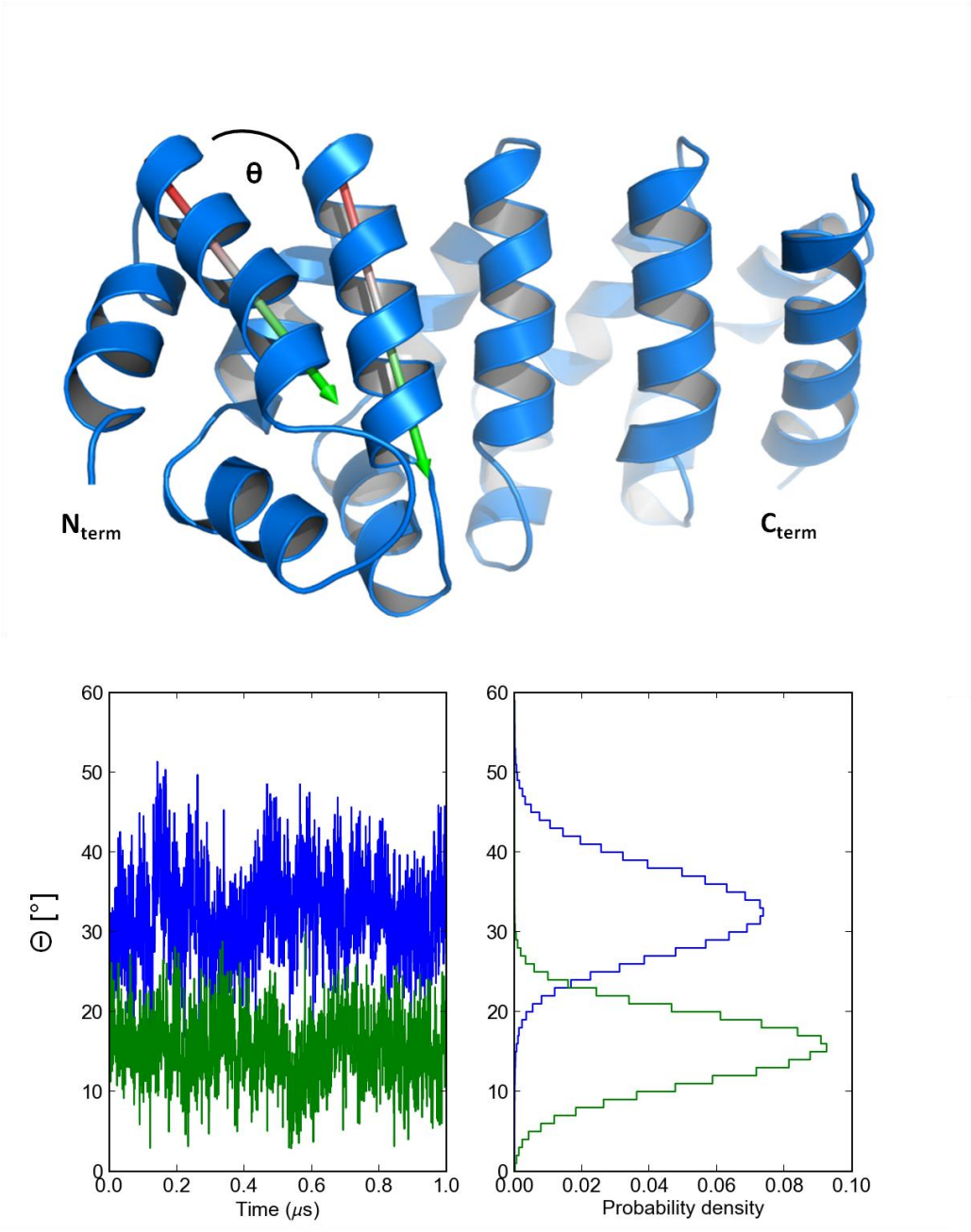


Figure 5.10 Comparison of $Y_IMR^1R^2A_{II}$ and $Y_{III}M_3A_{III}$ analyzing the rotational movement of the N-cap of $Y_IMR^1R^2A_{II}$. The angle θ between the two vectors representing helix 2 of the N-cap and helix 3 of the first internal repeat depicted on the top, was tracked over the course of the trajectory for both proteins. Bottom: $Y_IMR^1R^2A_{II}$ (blue) displays significantly higher flexibility (average θ 32.7 ± 5.5°) than $Y_{III}M_3A_{III}$ (green, average θ 15.6 ± 4.3°).

5.4.6. Binding Strengths of NT Towards Various N-cap Mutants of $Y_IMR^1R^2A_{II}$ as Probed by NMR

Obviously, the exact nature of the N-cap has a dramatic effect on the capability of the proteins to bind NT, and the stabilized mutants of the original binder VG_328 did not bind NT any longer. Therefore, the effect of the differences between the N_I- and the N_{II}-cap on peptide binding was investigated in detail using a series of mutants. The complete change from a N_I-cap to a N_{II}-cap includes three mutations: V34R, R37S and R42Δ. All single and possible double mutations as well as the triple mutation (= N_{II}-cap) were introduced into $Y_IMR^1R^2A_{II}$. Moreover, R42A was introduced as single point mutation to distinguish between the effect of shortening the loop between the N-cap and the first internal repeat, as done with R42Δ, and removing the R42 side chain as a potential point of interaction with NT. Furthermore, E46A was introduced to probe for the effect of removing a negatively charged residue from said loop as NT contains several positively charged residues.

The original full-length binder $Y_IMR^1R^2R^3MA_I$ (VG_328), its binding-competent version with stabilized C-cap $Y_IMR^1R^2R^3MA_{II}$, and the stablized but binding-incompetent N-cap variant $Y_{II}MR^1R^2R^3MA_{II}$ were made to serve as reference proteins. Additionally, VG_306 a version of VG_328 with the single point mutation Y116H was included in the study. VG_306 has been previously shown to display only about 50 % of the original signal intensity for NT binding in ELISA studies, pointing to an important role of Y116 for NT binding [24]. CSPs from the titration of ^{15}N-labeled protein with NT were used to determine the K_d values of all protein variants as described under Materials and Methods. In addition we titrated $Y_IMR^1R^2A_{II}$ with NT7-13, a truncated version of NT. All results are summarized in Figure 5.11and Table 5.1. Exemplary fitted raw data for $Y_IMR^1R^2A_{II}$ are shown in Figure 5.12.

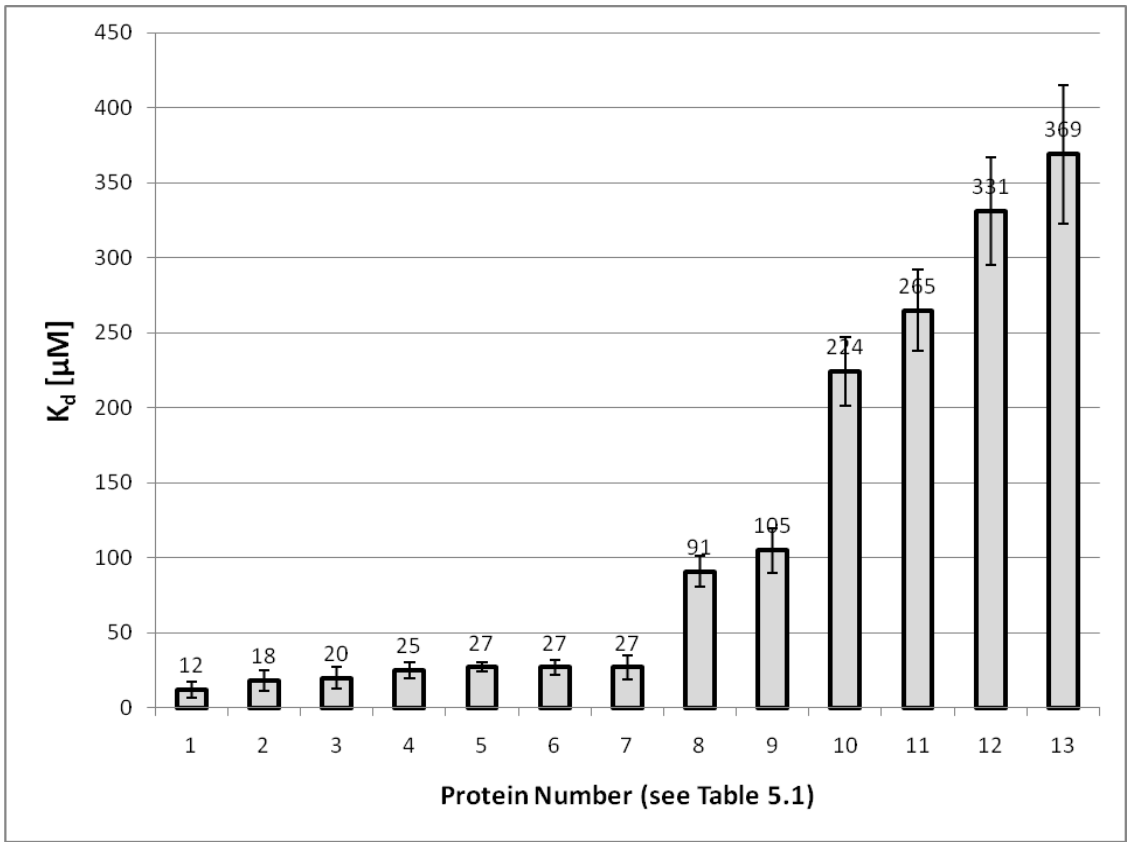


Figure 5.11 Dissociation constants (K_d) as determined by NMR using CSPs.

Table 5.1 List of dissociation constants of $Y_IMR^1R^2A_{II}$ mutants and VG_328 based reference proteins. For each protein the CSPs of several residues (n) were tracked, fitted, and the obtained K_d values averaged.

No.	Protein	Peptide	K_d [µM]	Relative K_d (relative to $Y_IMR^1R^2A_{II}$)	K_d represents Ø of n curves
1	$Y_IMR^1R^2A_{II}$	NT7-13	12±5	0.7	13
2	$Y_IMR^1R^2A_{II}$	NT	18±7	1.0	11
3	$Y_IMR^1R^2R^3MA_I$ (VG_328)	NT	20±7	1.1	6
4	$Y_IMR^1R^2A_{II}$_R42A	NT	25±5	1.4	9
5	$Y_IMR^1R^2A_{II}$_V34R	NT	27±3	1.5	7
6	$Y_IMR^1R^2A_{II}$_E46A	NT	27±5	1.5	7
7	$Y_IMR^1R^2R^3MA_{II}$	NT	27±8	1.5	7
8	$Y_IMR^1R^2A_{II}$_R42Δ	NT	91±10	5.1	13
9	$Y_IMR^1R^2A_{II}$_V34R_R42Δ	NT	105±15	5.8	11
10	$Y_IMR^1R^2A_{II}$_R37S	NT	224±23	12.5	12
11	$Y_IMR^1R^2A_{II}$_V34R_R37S	NT	265±27	14.7	6
12	$Y_IMR^1R^2A_{II}$_R37S_R42Δ	NT	331±36	18.4	8
13	$Y_{II}MR^1R^2A_{II}$	NT	369±46	20.5	7

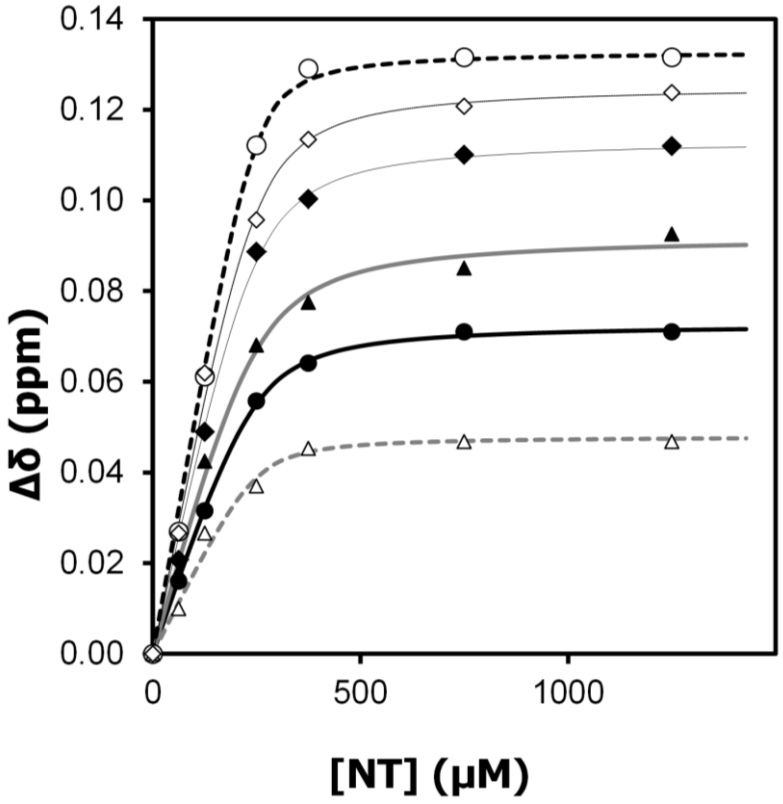


Figure 5.12 Fitted CSP raw data of exemplary $Y_IMR^1R^2A_{II}$ residues; G86 (empty circle), G128 (empty diamond), L76 (filled diamond), S66 (filled triangle), A59 (filled circle), A92 (empty triangle).

Interestingly, the titration of $Y_IMR^1R^2A_{II}$ with NT7-13 (see Figure 5.9, bottom peptide 6) yielded a K_d of 12±5 µM in the same range as the full-length peptide, indicating that the first 6 residues of NT do not contribute to binding. This view is supported by the diffusive and weak attenuation pattern of the N-terminal PRE-tag described above (see Figure 5.9, top left Q1C-MTSL).

The K_d of $Y_IMR^1R^2A_{II}$ was 18±7 µM, the V34R mutation was found to have the smallest impact on binding, with the K_d remaining in the same range as observed for $Y_IMR^1R^2A_{II}$. R42Δ increased the K_d significantly by a factor of 5, while R37S was found to be the most disruptive point mutation increasing the K_d by a factor of 12.5. Combinations of the single mutations showed synergistic effects, e.g. V34R/R37S with a factor of 15 and V34R/R42Δ with a factor 6. The effect of all three mutations present in $Y_{II}MR^1R^2A_{II}$ led to an increase in K_d by a factor 20.5 to 369±46 µM. The K_d of the original binder VG_328 ($Y_IMR^1R^2R^3MA_I$) was found to be 27 µM and the stabilized binder $Y_IMR^1R^2R^3MA_{II}$ yielded a K_d of 31 µM. In comparison the optimized minimal binder $Y_IMR^1R^2A_{II}$ has at least equal affinity for NT.

5.4.7. Binding Strengths of NT Towards Various N-cap Mutants of $Y_IMR_1R_2A_{II}$ Confirmed by ELISA and SPR Studies

Additionally, all $Y_IMR^1R^2A_{II}$ variants and VG_328-based reference proteins described above were assessed for interaction with NT in ELISA studies to verify NMR derived binding data (see Figure 5.13). Biotinylated NT was used as coated target as described under Materials and Methods. ELISA results of the $Y_IMR^1R^2A_{II}$ variants and the VG_328 derived reference proteins were in good agreement with CSP-based K_d results confirming the trends described above. We found that $Y_IMR^1R^2A_{II}$ gave a significantly higher signal than the original VG_328 and $Y_IMR^1R^2R^3MA_{II}$ indicating a size-dependent effect between the surface bound ELISA method and in solution NMR method. For proteins of the same size a similar intensity of the ELISA signal translated reliably into a range of similar K_d values in NMR experiments.

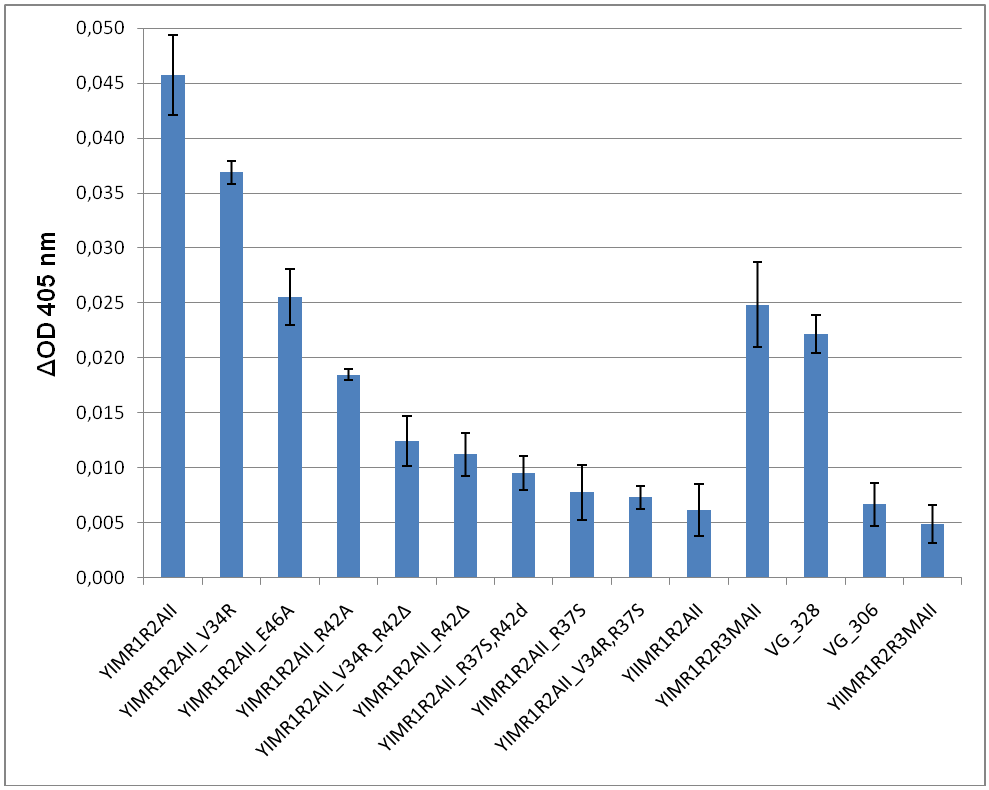


Figure 5.13 Affinity of $Y_IMR^1R^2A_{II}$ mutants and VG_328 based reference proteins for NT1-13 detected by ELISA. 200 µM protein, 200 µM target peptide NT_{biotin} (details see Table 5.5), read out at 405 nm after 2 hr, reference wavelength and unspecific peptide binding background deducted (see Materials and Methods).

Furthermore, the affinity of $Y_IMR^1R^2A_{II}$ and $Y_IMR^1R^2R^3MA_{II}$ for NT was determined by SPR as described under materials and methods. At 8°C we determined 14 µM for $Y_IMR^1R^2A_{II}$ and 18 µM for $Y_IMR^1R^2R^3MA_{II}$ (see Figure 5.14). Earlier studies determined a K_d of 7 µM for the original binder VG_328 ($Y_IMR^1R^2R^3MA_I$) with the old A_I-type C-cap at 4 °C with a similar experimental set-up [24]. The much lower temperatures in comparison to NMR experiments (37 °C) account for the lower K_d values obtained by SPR. It should be noted that for binders in the µM range K_d values from SPR have higher errors compared to those obtained from NMR experiments, as NMR is better suited to accurately determine affinity constants in this range.

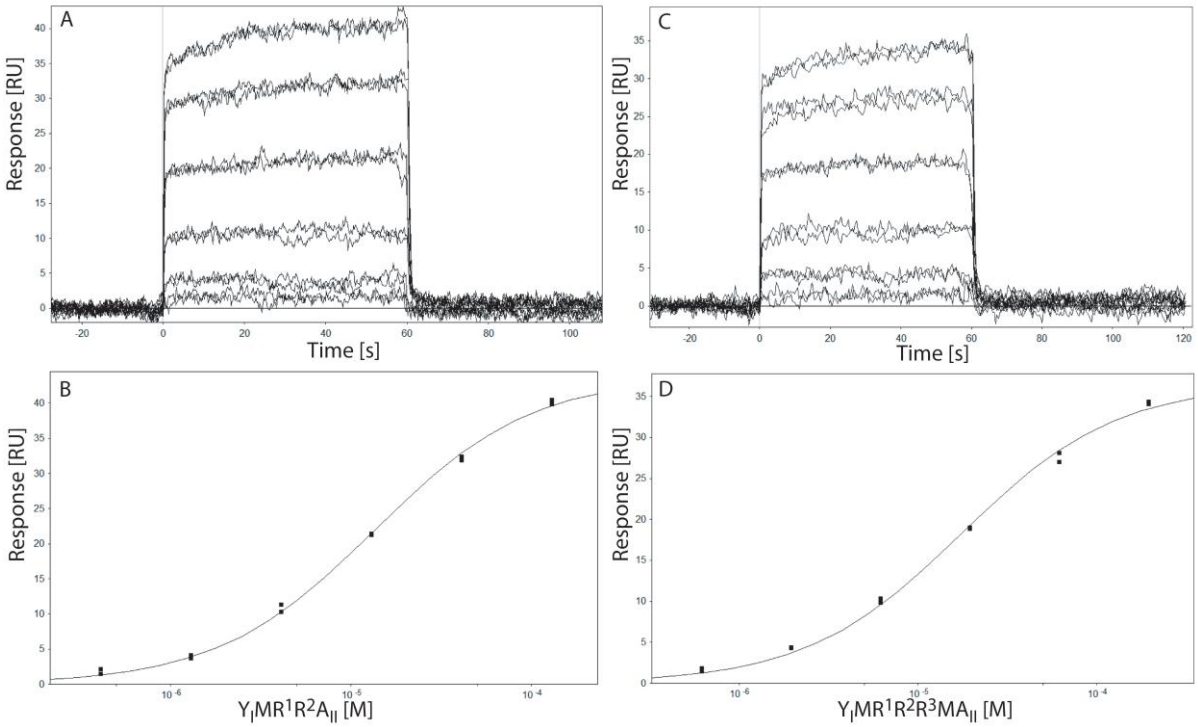


Figure 5.14 SPR response curves (top) and fitted data (bottom) for $Y_IMR^1R^2A_{II}$ (A, B), 14 µM) and $Y_IMR^1R^2R^3MA_{II}$ (C, D, 18 µM) at 8 °C.

5.4.8. Probing Protein Stability with MD Simulation

To obtain insight into the effects of different mutations on protein stability MD simulations of $Y_IMR^1R^2A_{II}$ and its N-cap variants without peptide were carried out in explicit water. By monitoring root mean square fluctuations (RMSF) over the course of the simulation we could confirm that the effects of the N-cap mutations were restricted to the N-terminal part of the protein (data not shown). $Y_IMR^1R^2A_{II}$ and $Y_IMR^1R^2A_{II}$_V34R display a similar pattern of fluctuations, with the highest fluctuations in the loop region between the N-cap and the first repeat (residues 41-46) (see Figure 5.15). Both $Y_IMR^1R^2A_{II}$ and $Y_IMR^1R^2A_{II}$_V34R were stabilized by the presence of NT, with $Y_IMR^1R^2A_{II}$_V34R showing a stronger stabilization than $Y_IMR^1R^2A_{II}$ in complex with NT. This difference was most pronounced for helix 2 of the N-cap and the adjacent loop between the N-cap and the first repeat, as well as for helix 3 of the first repeat. Simulations of the R42Δ deletion showed highly similar effects to the V34R mutation on the RMSF, indicating increased flexibility in the loop connecting the N-cap and the first internal repeat (data not shown).

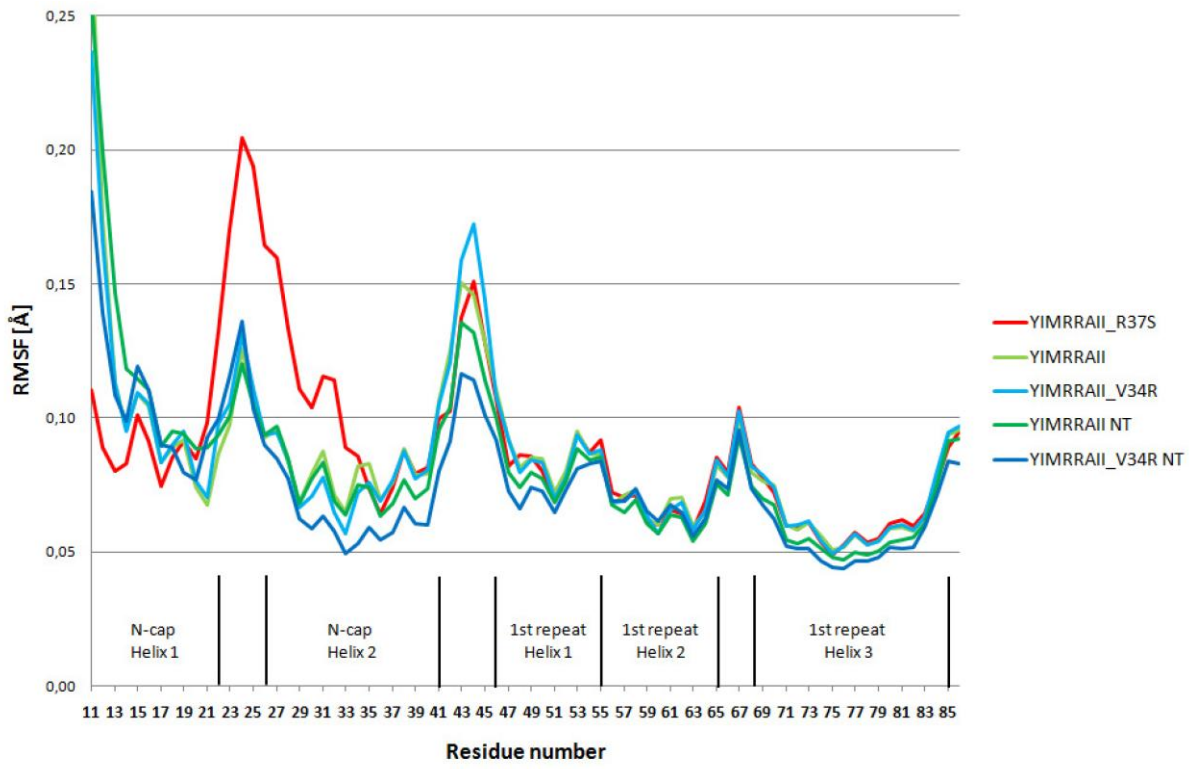


Figure 5.15 Comparison of RMSF values for the N-cap and the first internal repeat of selected $Y_IMR^1R^2A_{II}$ variants in presence and absence of NT. The location of helices is indicated by vertical bars. No binding of NT could be detected for $Y_IMR^1R^2A_{II}$_R37S.

A key stabilising feature of the N-cap appears to be the salt bridge formation between R37 in the N-cap and E74 in the first internal repeat. In $Y_IMR^1R^2A_{II}$ and $Y_IMR^1R^2A_{II}$_V34R this salt bridge is reliably formed during extended periods of time in the simulation (see Figure 5.16 D). Only for a small fraction of time an alternative salt bridge occurs between R34 and E74 in $Y_IMR^1R^2A_{II}$_V34R (see Figure 5.16 A), while for most of the time R34 is involved in a hydrogen bond with the side chain oxygen of Q70 (see Figure 5.16 B). The side chain of R34 is further stabilized by π-π- and electrostatic interaction with the side chain of R37 (see Figure 5.16 C).

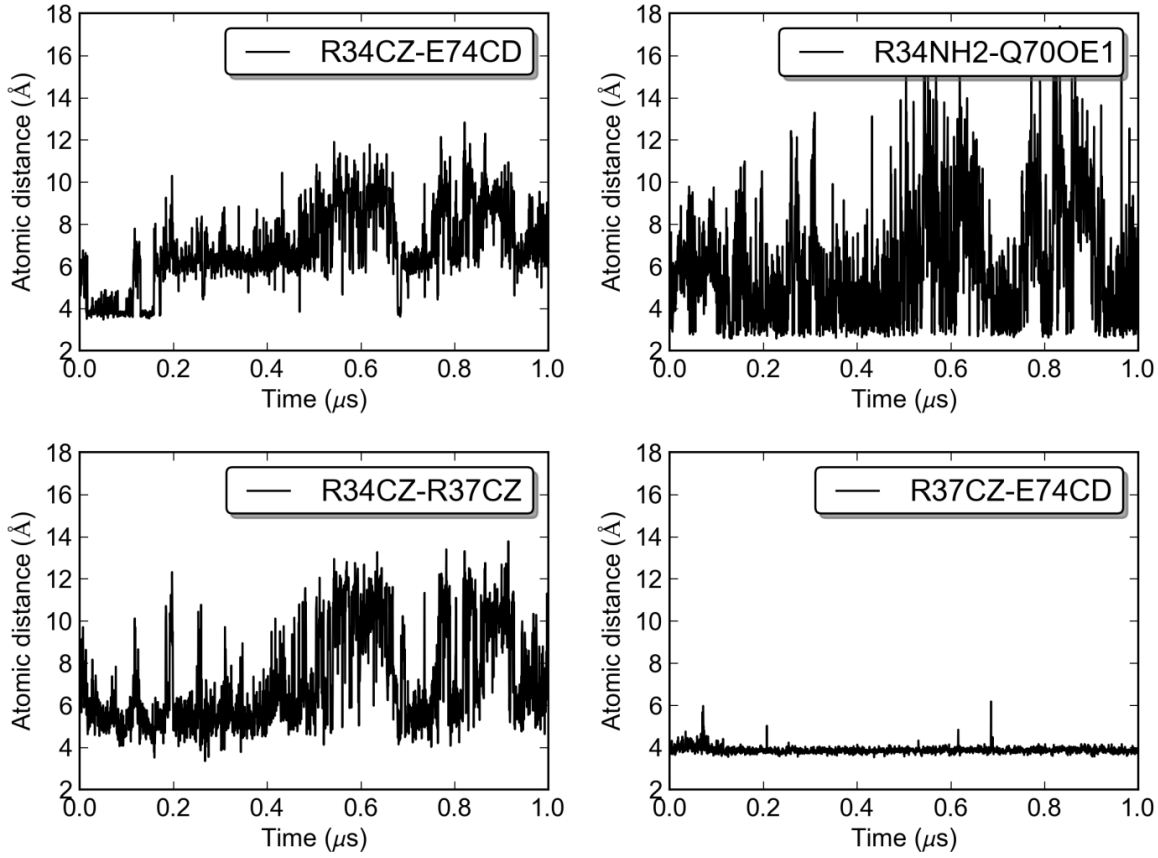


Figure 5.16 Evolution of atomic distances in $Y_IMR^1R^2A_{II}$_V34R over the course of the trajectory. In a small portion of the simulation time, a salt bridge between R34 and E74 was observed (top-left). A hydrogen bond between R34 and the sidechain oxygen of Q70 was observed in most of the simulation time (top-right). Moreover, the sidechain of R34 forms π-π- and electrostatic interactions with the aliphatic part of the sidechain of R37 (bottom-left). E74 predominantly forms a very strong salt bridge with R37 (bottom-right).

Accordingly, the elimination of R37 in $Y_IMR^1R^2A_{II}$_R37S causes the largest fluctuations observed in the series, noticeably destabilizing the fold of the N-cap (see Figure 5.15). In this mutant no other cationic residue is present in the vicinity that could contribute to the salt bridge, which helps to stabilize packing of the N-cap against the first internal repeat. In contrast, R34 replaces the function of R37 in the double mutant $Y_IMR^1R^2A_{II}$_V34R_R37S by forming a salt bridge with E74 (see Figure 5.17 A). The simulations also indicate the presence of a weak hydrogen bond between S37 and E74 (see Figure 5.17 B). The fully stabilized protein $Y_{II}MR^1R^2A_{II}$ settled into a stable fold after 200 ns and retained it for the remainder of the simulation, remarkably, $Y_IMR^1R^2A_{II}$_V34R in the presence of NT still showed superior

stability to $Y_{II}MR^1R^2A_{II}$ (data not shown). Further simulations of the effects of combinations of the single mutations are still in progress and will be evaluated as soon as the results of these simulations become available.

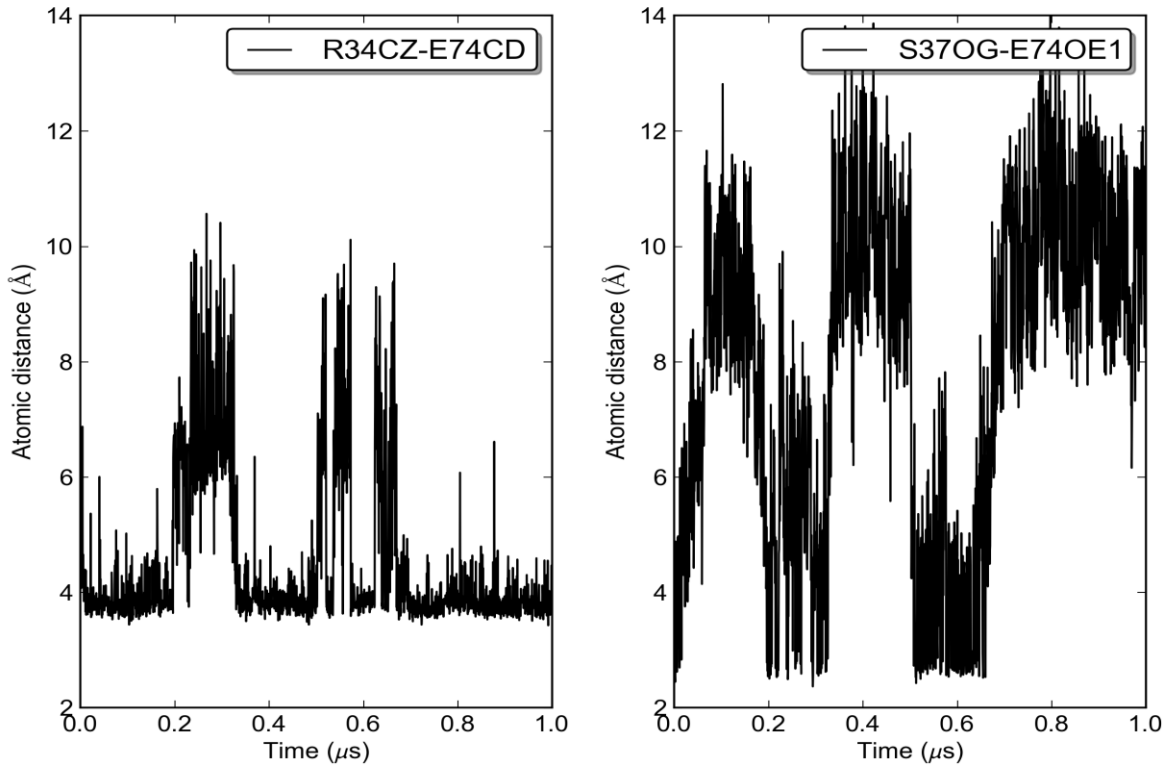


Figure 5.17 Evolution of atomic distances in $Y_IMR^1R^2A_{II}$_V34R_R37S during the MD trajectory. E74 predominantely forms a very strong salt bridge with R34 (left). A weak hydrogen bond between S37 and E74 was observed for short amounts of time (right).

5.4.9. Identifying Potential Binding Conformations of NT Using MD Simulations of $Y_IMR^1R^2A_{II}$ and its Variants in Complex with NT

Data from CSP, PRE and mutational studies on protein-NT interactions were insufficient to unambiguously place NT in its bound position. Moreover, it was unclear whether NT binds in a unique mode. To obtain more insight into the details of binding we calculated extended MD trajectories of the $Y_IMR^1R^2A_{II}$-NT complex. In these calculations NT was placed in different starting conformations that could potentially satisfy all the CSP, PRE and mutation data, and 2-4 μs simulations of the $Y_IMR^1R^2A_{II}$-NT complex were computed as described in Materials and Methods. In all these trajectories the C-terminal part of NT (P7-L13) assumed a stable horizontal and anti-parallel conformation along the upper half of the binding surface formed by the helices 3 during long periods in the simulations (see Figure 5.18 and Figure 5.19). This

location satisfied all experimentally identified contact points and involved numerous randomized surface residues of $Y_IMR^1R^2A_{II}$.

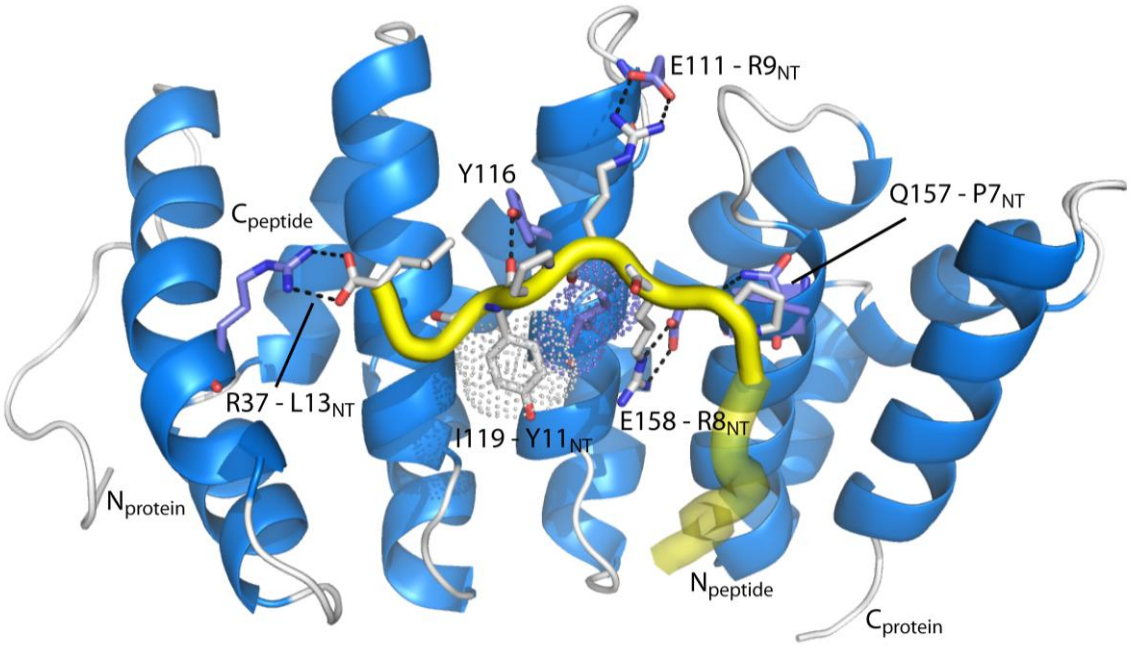


Figure 5.18 Peptide conformation picked at 19.465 µs. $Y_IMR^1R^2A_{II}$ is displayed in blue, NT in yellow. Side chains of interacting residues are depicted as sticks. Dotted lines indicate observed interactions. The N-terminal part of NT is displayed transparently to indicate that this part of the conformation is less stable over the course of the simulation. The hydrophobic interaction between Y11 of NT and I119 of the protein is depicted by a dotted surface. Peptide residues involved in interactions are labeled as X_{NT}. A detailed description of the interaction is given in the text.

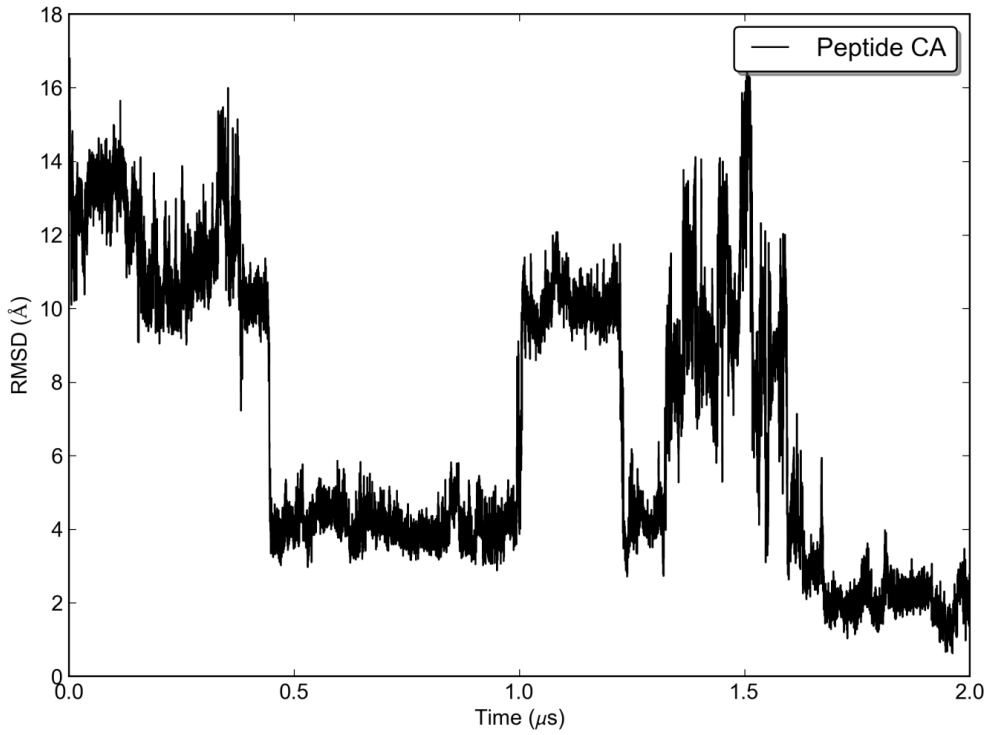


Figure 5.19 RMSD fluctuations of the peptide Cα-atoms over the trajectory in comparison to the reference peptide conformation picked at 19.465 µs (protein conformations were superimposed). The RMSD plot indicates that the peptide has sampled a large amount of the conformational space, and can remain in this conformation for a relatively a long time (approx. 0.5 µs - 1.0 µs and 1.6 µs - 2.0 µs). The starting conformation of the peptide in this simulation was vertical to the binding surface, aligned with helix 3 of the fist internal repeat. The resulting stable conformation found was horizontal across the binding surface and is shown in Figure 5.18.

We observed that all four side chains of the peptide (P7, R8, R9 and Y11), which had been identified as crucial to peptide binding in the Ala-scan [24], make strong interactions with the protein. The backbone carbonyl of P7 forms a hydrogen bond with the side chain amide moiety of Q157 in helix 3 of the third internal repeat (R^2). The guanidinium group of R8 forms a salt bridge with the side chain carboxylate of E158 (randomized position) in the same repeat. Furthermore, the guanidinium group of R8 is located to allow π-stacking with Trp161 (randomized position). A second salt bridge is formed between R9 and E111 (randomized position) in helix 3 of the second internal repeat (R^1). Furthermore, the backbone amide of R9 makes a hydrogen bond with the carboxylate of Q115. Finally, hydrophobic contacts are made between Y11 and I119 (randomized position) and with the hydrophobic part of the side chain

of R8 and between L13 and W77. In addition, we observed that the carboxyl group of the free peptide C-terminus forms a salt bridge with R37 in the N-cap, explaining why the R37S mutation reduces binding affinity.

On the protein side the hydroxyl of Y116 forms a hydrogen bond with the backbone carbonyl of P10. Previous studies have shown a significant drop in binding to NT when Y116 is mutated to His [24]. The N-terminal part of the peptide (Q1-K6) alternated between several conformations, including complete detachment and the formation of a hairpin that places N- and C-terminus in vicinity. These dynamics are supported by the PRE data from the NT variant, in which MTSL was attached to the N-terminus, and in which the attenuations were weak and spread over a large area (*vide supra*).

For further verification we analyzed the observed peptide-binding mode of $Y_IMR^1R^2A_{II}$ and NT for agreement with PRE-data. This was achieved by adding distance restraints representing the conformational restraints of the spin label and its effective range based on the positions sampled by the respective side chain during the course of the simulation. We chose 5 Å as effective range in which we expected strong signal attenuation. Preliminary analysis of the data indicates a reasonable overlap of protein regions attenuated *in vitro* and regions within the simulated sphere of the spin label effect *in silico*.

5.5. Discussion and Conclusions

The present study clearly demonstrates the enormous difficulties faced when analyzing weak protein-peptide interactions. While the project has shown how difficult it is to obtain data on the binding mode, the results undoubtedly helped drive forward the design and construction of engineered ArmRPs.

Clearly, the bottleneck in studying peptide-protein interactions by NMR is the requirement for assignments, and these are particularly difficult to obtain for repeat proteins. Another problem at the onset of the project was the limited stability of the protein, which, however, was rapidly improved through introduction of stabilized N-caps (V24R, R27S and R32Δ = Y_{II}) and C-caps (Q292L and F293Q = A_{II}) as described by Alfarano *et al.*[26]. However, when the N-cap mutations were transferred to the randomized binder ($Y_{II}MR^1R^2R^3MA_{II}$), binding of NT was severely affected. Nevertheless, the incorporation of the type A_{II}-cap alone improved the stability of VG_328 ($Y_IMR^1R^2R^3MA_I$) such that the NMR sample was stable over months at 310 K. Unfortunately, the size of $Y_IMR^1R^2R^3MA_{II}$, 32 kDa, in combination with the highly repetitive sequence did not allow extensive assignments.

It was noted in other work (Watson *et al.*, in preparation) that N-terminally truncated armadillo repeat proteins resulted in [^{15}N,^{1}H]-HSQC spectra that were to a large degree super-imposable with spectra of their full length parent. In contrast, C-terminally truncated fragments displayed molten globule-like behavior regardless of their length. This points to a crucial role of the A_{II}-cap for protein stability, an effect that is transferred through the complete protein. Apparently, the stability of the Y_I-cap is by far inferior, a view that is supported by our observation that many signals from the N-cap were missing in the [^{15}N,^{1}H]-HSQC spectra. We noticed a similarly decisive role of cap stability for overall protein stability before in the class of Ankyrin repeat proteins [29]. Additionally, in the series of N-terminally truncated proteins, we noticed that $R^1R^2R^3MA_{II}$, as opposed to other proteins of this series, displays severe line-broadening indicative of the formation of oligomers akin to a larger hydrophobic interface being solvent-exposed for R_1. Taken together these data provide valuable insight into the individual contributions of caps and internal repeats for overall protein stability. While the systematic truncation of N-terminal repeats from $Y_IMR^1R^2R^3MA_{II}$ by one repeat at a time allowed us to extend the backbone resonance assignments we were

still unable to fully assign $Y_IMR^1R^2R^3MA_{II}$ and its complex with NT. Nevertheless, the assignments obtained that way allowed analysis of CSP data to an extent which demonstrated that NT binding to $Y_IMR^1R^2R^3MA_{II}$ had no effect on the resonances of the C-cap and the last internal repeat.

As a consequence of knowing approximately where NT was bound we could eliminate whole repeats that we suspected to be unimportant for peptide binding, while retaining the stabilizing caps. The resulting optimized minimal binder $Y_IMR^1R^2A_{II}$ retained NT binding properties and displayed improved spectra. As $Y_IMR^1R^2A_{II}$ represents a much smaller target – 22 kDa instead of 32 kDa – and because assignments were easier in the absence of two identical unrandomized M-repeats, we could achieve near-complete and unambiguous backbone assignment of the whole protein and partial side-chain assignments for the binding interface formed by helix 3 of each repeat. Thereby, detailed residue-by-residue analysis of the interaction with NT by CSP and evaluation of PREs became possible. This strategy of pruning unnecessary repeat modules from established binders at no expense in binding affinity represents a significant, and hereto unutilized, advancement in the development process of repeat protein engineering.

CSP experiments of $Y_IMR^1R^2A_{II}$ with NT revealed substantial changes in the binding interface and the N-cap, and smaller changes in the hinge regions between helices. The effects were spread over a much larger area than expected. We interpreted this as a combination of direct effects due to peptide binding and indirect effects due to structural rearrangements involving the N-cap. We hypothesize that the N-cap is locked into one position upon binding of NT causing a series of strong CSPs in the interface between the N-cap and the first internal repeat. These stabilizing effects are propagated from the cap through the whole protein leading to minor CSPs in hinge regions. The strong CSPs on the binding surface of helices 3 are most likely due directly to NT binding. As the low affinity did not permit detection of intermolecular NOEs and the attainment of a more precise picture of the binding mode we resorted to the more sensitive PRE method by attaching spin labels to the peptide. The unaltered interaction of NT7-13 with $Y_IMR^1R^2A_{II}$ in NMR studies indicated that the N-terminal hexapeptide of NT does not contribute significantly to binding, an observation that was confirmed by the diffuse pattern observed in PRE studies (see Figure 5.9). PRE studies also showed that the central region of the peptide is more rigidly located in the complex, whereas the C-terminus samples a number of conformations in non-continuous regions of the

protein surface. The most notable contact identified in MD simulations was a salt bridge formed by the carboxylate of the free C-terminus of the peptide and R37 from the N-cap. Considering that Y11, the last residue of importance for side chain interactions in the Ala-scan (*vide infra*), is still two residues distant from the C-terminal PRE-tag, the possible motion of the PRE-tag may exert its quenching effect over a larger area whenever this salt bridge is not formed providing a rationale for the diffuse PRE pattern observed for the C-terminal spin label. This effect could be even stronger when a large moiety like a spin label is attached to the C-terminus leading to more mobility in comparison to the unmodified peptide used for the CSP experiments and the MD simulations. The observed range of PRE effects was confirmed by MD simulations (*vide infra*).

An unexpected observation mentioned above, was that the protein mutants incorporating the stabilized N-caps no longer bound NT. Initially it was unclear whether this effect was due to removal of residues that form contacts with NT, or whether the geometry of the binding interface was altered and incompatible with NT binding.

The mechanism by which this binding is affected is of considerable interest. Interestingly, the three mutations V34R, R37S and R42Δ, that change the binding competent into the binding incompetent N-cap, all affect positioning of arginine residues. We studied the mutations individually to determine which of them were responsible for increased protein stability and which were crucial for NT binding. Titrations of NT against $Y_IMR^1R^2A_{II}$ and its N-cap variants enabled us to identify R37S as the most disruptive single mutation for peptide binding increasing the K_d from ~18 µM to about 224 µM (factor 12.5). In contrast, the V34R mutation had almost no influence, whereas R42Δ increased the K_d by a factor of 5 (See Figure 5.11 and Table 5.1). Combinations of the single mutations showed synergistic effects, with the triple mutant $Y_{II}MR^1R^2A_{II}$ displaying a 20.5-fold K_d.

Finally, we established an approximate binding model of the $Y_IMR^1R^2A_{II}$: NT complex using unrestrained molecular dynamics simulations. The outcome of the simulation was compatible with all experimental data, and can therefore be seen as an unbiased confirmation of the experimental results. The observed salt bridge between R37 in the N-cap and the free C-terminus of the peptide helps to explain why so far all stabilized cap designs containing the R37S mutation had failed to retain peptide binding. The mutation V34R was confirmed to introduce an alternative contact to E74 in the first internal repeat, providing a cap stabilizing effect, while still allowing R37 to form a contact with the peptide. In conclusion we have

identified that $Y_IMR^1R^2A_{II}$_V34R binds NT with highly similar affinity and manner to $Y_IMR^1R^2A_{II}$ albeit with a more stable fold in the complex, thereby presenting an ideal compromise between affinity and stability.

The evidence presented here suggests that the central part of the NT peptide is bound to the upper part of the designed binding interface across helices 3 of all three internal repeats involving numerous randomized protein positions. However, it appears that NT is not tightly bound in a unique extended conformation and does not utilise the conserved asparagine ladder for backbone binding. This is not entirely surprising as the peptide target is proportionally too long for the protein scaffold at hand. Theoretically an ArmRP with three internal repeats should accommodate a hexapeptide, whereas NT has 13 residues. When one examines the sequence of the peptide, one can see that the two proline residues reduce the ability of the peptide to easily assume a fully extended conformation. This supports our finding that the central part of the peptide is bound more tightly and that the N-terminus makes only transient interactions with other parts of the $Y_IMR^1R^2A_{II}$ binding interface. The binding hypothesis from this work is supported by the results from the Ala-scan of NT [24], which indicated that a cluster of residues in the central part of the 13-residue peptide NT (Residues 7-9 and 11) is important for binding. It is also in agreement with the fact that the binder was developed by pre-panning against the first 5 residues of NT during the ribosome display selections.

In conclusion we could demonstrate that the central part of NT (residues 7 to 11) makes contacts with the binding interface presented by helices 3 of $Y_IMR^1R^2A_{II}$ as intended in the original design. Moreover, NT is bound in anti-parallel fashion as observed for peptide ligands in naturally occurring ArmRPs. Additionally the C-terminus of the peptide forms a strong salt bridge with R37 in the N-cap, further improving binding affinity. We successfully reduced the size of the original binder from 32 to 22 kDa without loss of binding competency, and confirmed the NT sequence PRRPYIL as the key part of the peptide for binding.

For future work one could aim to transfer the binding pattern of NT onto an ArmRP scaffold with a stabilized N-cap. An additional M-type repeat with an arginine in a suitable position in helix 3 could be inserted preceding the established repeat modules of $Y_IMR^1R^2A_{II}$ creating $Y_{II}M_{arg}MR^1R^2A_{II}$ abolishing the need for the relatively unstable first generation N_I-cap.

Undoubtedly, crystal structures of larger proteins provide fast access to structural information of protein-peptide complexes, and this particularly true for repeat proteins. However, in early

stages of such projects binding affinities are low, and protein binders may still contain flexible parts hampering crystallization. Herein we have developed a highly interdisciplinary approach combining mutagenesis, heteronuclear NMR spectroscopy and MD calculations as well as other biophysical tools. We believe that this approach can be a powerful strategy for analyzing difficult targets such as low affinity binders with multiple conformations and limited stability. We have also demonstrated that, even with information from limited NMR assignments, protein sequences can be modified to yield proteins with superior characteristics that may eventually be amenable to high-resolution structural studies. Most importantly, this limited information is sufficient to drive the project forward and to verify original hypotheses about the binding mode of the ligand in designed binders. In particular in the early stages of such projects it is of the utmost importance to ensure that the project is on the correct track.

5.6. Supplementary Materials

Table 5.2 Oligonucleotides

Name	Sequence 5'-3' direction	Purpose
078_RMAm_FWD	GAA AAT TTA TAT TTT CAG GGT AAC GAA CAA CAA	Construction of R^3MA_{II} from pPANK_N_IM328MC_{II}
077_R23MAm_FWD	GAA AAT TTA TAT TTT CAG GGT AAC GAA CAA ACC	Construction of $R^2R^3MA_{II}$ and $R^1R^2R^3MA_{II}$ from pPANK_N_IM328MC_{II}
090_MRRRMAm_v10_FWD	GAA AAT TTA TAT TTT CAG GGT CGT GAT GGT AAC GAA CAA	Construction of $MR^1R^2R^3MA_{II}$ from pPANK_N_IM328MC_{II}
088_RxMAm_v3_REV	AGA TGA GAG TAA GGC TAT CAT TAG TGG GAC TGC AG	Reverse primer for 077 ($R^2R^3MA_{II}$), 078 (R^3MA_{II}) and 090 ($MR^1R^2R^3MA_{II}$)
053_B328_YMR_x_pPANK_Fwd	TAA TGA GGT ACC CCG GGT CGA CCT GCA GCC	Construction of $Y_IMR^1R^2R^3$ and Y_IMR^1 from pPANK_NyM328MCa encoding VG_328
054_B328_YMRRR_pPANK_REV	CAG TTC AGC GAT GTT AGT CAG AGC GTC CAG	Reverse primer for 053 ($Y_IMR^1R^2R^3$) and 101 ($Y_IMR^1R^2R^3A_{II}$)
056_B328_YMR_pPANK_REV	AGC GAA AGC GAT GTT GTT CAG AGC GAT AAG	Reverse primer for 053 (Y_IMR^1) and 101 ($Y_IMR^1A_{II}$)
063_YMRR_REV	TTC CAT AGC GAT GTT AGT C	Reverse primer for 65 ($Y_IMR^1R^2$) and 101 ($Y_IMR^1R^2A_{II}$)
065_YMRR_FWD	TAA TGA GGT ACC CCG GG	Construction of $Y_IMR^1R^2$ from pPANK_$Y_IMR^1R^2R^3$
101_AtypeC-cap_FW	GGT AAC GAA CAG AAA CAG GCT GTT AAA GAA G	Construction of $Y_IMR^1R^2R^3A_{II}$, $Y_IMR^1R^2A_{II}$, $Y_IMR^1A_{II}$ from pPANK_$Y_IMR^1R^2R^3MA_{II}$
118_delR42_FW	TGA TGG TAA CGA ACA AAT CC	Deletion of R42 from Y_I-cap
119_delR42_RV	GAC AGG ATC TGA CGG AAT TTA AC	Reverse primer for 118
121_R42A_FW	AGA TCC TGT CTG CTG ATG GTA ACG A	Introduction of R42A into Y_I-cap
122_R42A_RV	GAC GGA ATT TAA CGG TAG CAG	Reverse primer for 121
124_V34R_R37S_RV	AGC TGT TCC TGC ATG TCG TCG GAG TTC	Reverse primer for 139 and 140
139_Yf_V24R_FWD	GTC TGC TAC CCG TAA ATT CCG TCA GAT CCT G	Introduction of V34R into Y_I-cap and Y_I-cap_R42del
140_Yf_R27S_FWD	GTC TGC TAC CGT TAA ATT CTC TCA GAT CCT G	Introduction of R37S into Y_I-cap and Y_I-cap_R42del
141_M_E46A_FWD	GTC TCG TGA TGG TAA CGC ACA AAT CC	Introduction of E46A into Y_I-cap
142_M_E46A_REV	AGG ATC TGA CGG AAT TTA ACG GTA GC	Reverse primer for 141
152_V34R_FWD	GTC TGC TAC CCG TAA ATT CTC TCA G	Introduction of V34R into Y_I-cap_R37S and Y_I-cap_R37S_R42del
153_V34R_REV	AGC TGT TCC TGC ATG TCG TCG	Reverse primer for 152

Table 5.3 Minimal medium for isotopic labeling. For ^{2}H labeling salts were dissolved in D_2O, otherwise in H_2O.

^{15}N labeling	mM	**^{15}N,^{13}C labeling**	mM
K_2HPO_4	22.97	K_2HPO_4	22.97
KH_2PO_4	29.39	KH_2PO_4	29.39
$Na_2HPO_4 \cdot 2H_2O$	5.80	$Na_2HPO_4 \cdot 2H_2O$	5.80
$^{15}NH_4Cl$	9.35	$^{15}NH_4Cl$	9.35
D-(+)-glucose	25.23	^{13}C-D-(+)-glucose	10.09
Thiamine	0.15	Thiamine	0.15
$MgSO_4$	2.00	$MgSO_4$	2.00

Table 5.4 Trace metal solution for minimal media supplementation.

1000 × trace metal stock	g/L	mM
$FeSO_4 \cdot 7H_2O$	4	14.39
$CaCl_2 \cdot 2H_2O$	4	27.21
$AlCl_3 \cdot 6H_2O$	1	3.93
$MnSO_4 \cdot H_2O$	1	5.92
$CoCl_2 \cdot 6H_2O$	0.4	1.59
$ZnSO_4 \cdot 7H_2O$	0.2	0.70
$CuCl_2 \cdot 2H_2O$	0.1	0.68
H_3BO_4	0.1	1.28

Table 5.5 Peptides, free termini unless indicated otherwise. NT_{Biotin} composition: 6ACA (6-amino-caproic-acid), βA (β-alanine).

Peptide name	Sequence	Purpose
NT	pGlu-LYENKPRRPYIL	CSM and PRE-studies
NT7-13	Ac-PRRPYIL	CSM
NT_Q1C	CLYENKPRRPYIL	PRE-studies
NT_K6C	pGlu-LYENKPRRPYIC	PRE-studies
NT_L13C	pGlu-LYENCPRRPYIL	PRE-studies
NT_{biotin}	Biotin-6ACA-βA-βA-NT	ELISA

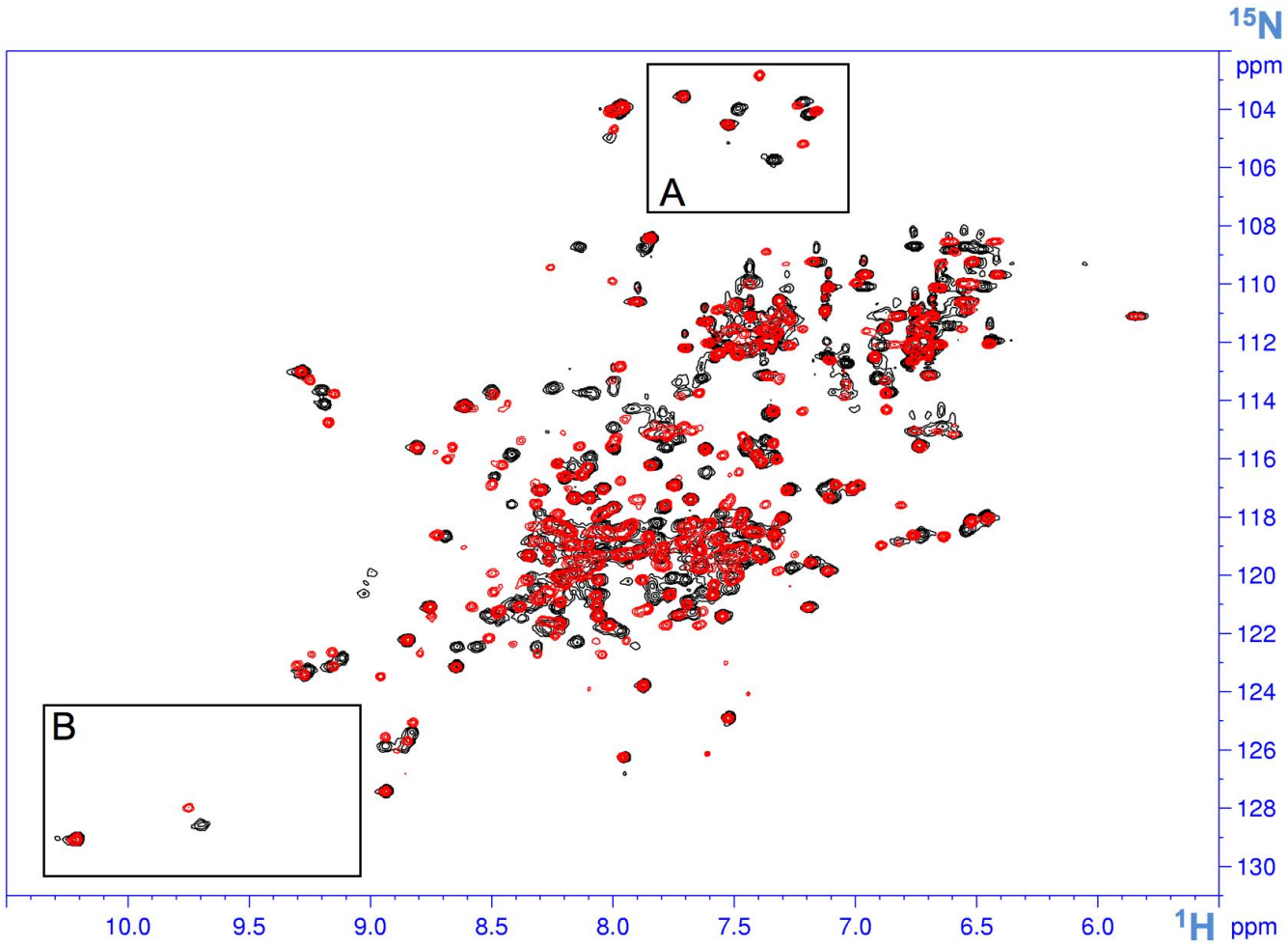


Figure 5.20 [^{15}N,^{1}H]-HSQC spectra showing chemical shift perturbations of 400 µM ^{15}N-labelled $Y_I MR^1R^2R^3MA_{II}$ without (black) and with (red) two equivalents of NT peptide. Details for the glycine (A)and tryptophane indole (B) regions are shown in Figure 5.7.

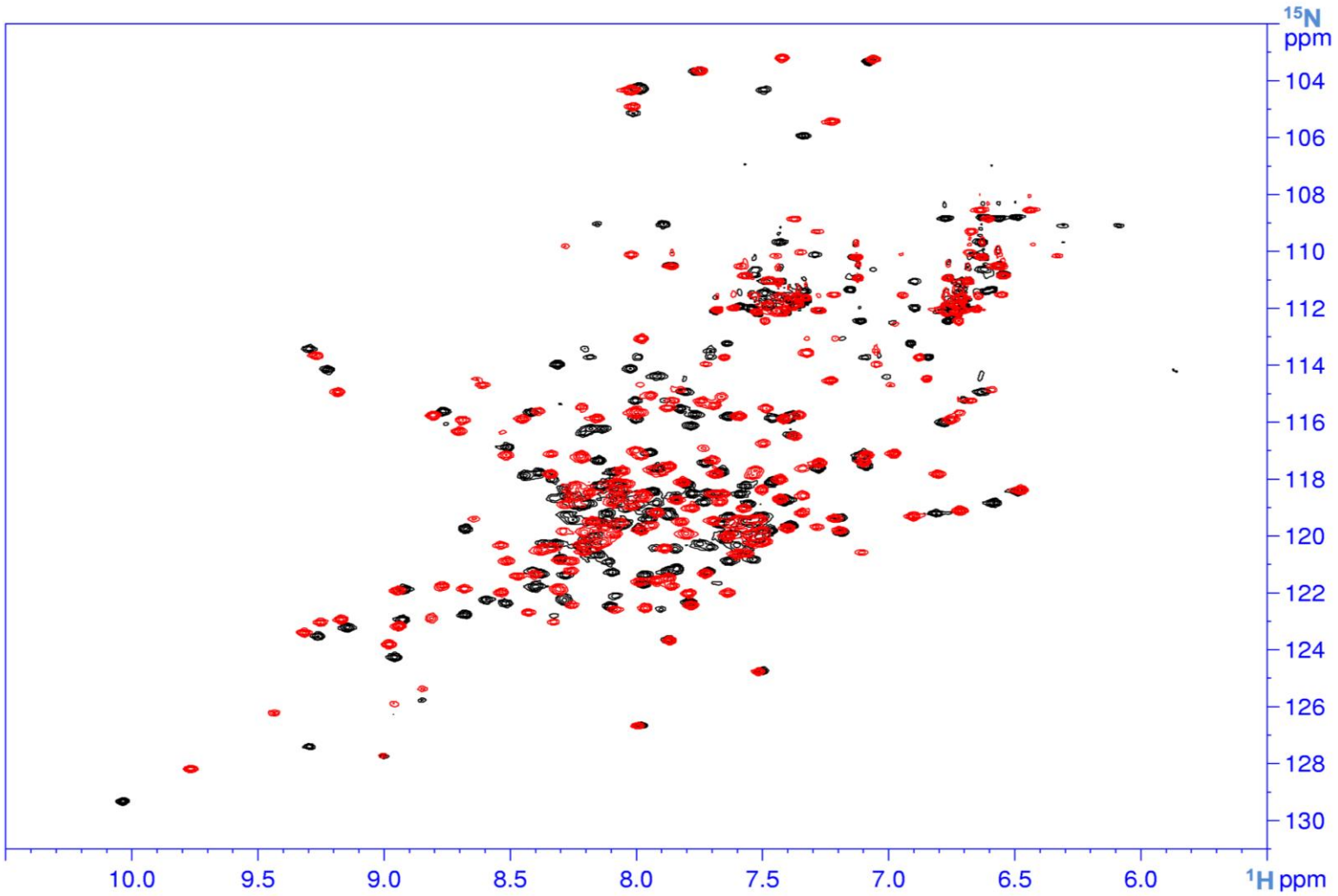


Figure 5.21 [^{15}N,^{1}H] HSQC spectra showing chemical shift perturbations of 400 µM ^{15}N-labelled $Y_IMR^1R^2A_{II}$ without (black) and with (red) two equivalents of NT peptide.

Table 5.6 List of chemical shifts for backbone and side chain assignments of ^{13}C, ^{15}N-labeled $Y_IMR^1R^2A_{II}$ in the presence of 2 equivalents NT.

Residue	H	N	C	CA	CB	CG*	CD*	CE*	HA*	HB*	HG*	HD*	HE*
PRO 15			179.5	66.5	30.8	29			4.42	2.36	1.81	3.73	
GLN 16	7.54	117.8	178.6	59.1	28.1	34.1			4.07	2.17	2.32		
MET 17	8.15	120.2	178.3	58.9	34.6	31.1			4.04	1.83	2.28		
THR 18	8.39	115.5		68	69.1	21.9					1		
GLN 19	7.5	120.1	179.5	58.9	28.2	33.6			4.04	2.17	2.44		
GLN 20	7.89	120.3	177.9	58.9	29	34.7			3.92	1.94	2.52/2.33		
LEU 21	7.5	116.7	177.2	57.4	42.3	27	23.7		3.96	2.01	1.27	0.76/0.60	
ASN 22	7.33	113.5	174.7	52.5	39.9				4.78	2.95/2.59			
SER 23	7.23	114.5	174.4	58.4	63.7				4.08	3.77			
ASP 24	8.33	122.9	175.6	54.8	40.8				4.47	2.63			
ASP 25	8.47	121.3	176.5	52.9	42				4.67	2.83/2.44			
MET 26	8.88	127.2	178.2	59.1	33.3	32.4		17.1	4.12	2.17/1.97	2.67/2.56		2.13
GLN 27	8.34	117.8	179.5	58.9	27.8	34.1			4	2.13	2.4		
GLU 28	7.64	121.9	178.9	59.1	29.3	36.7			4	2.13	2.32		
GLN 29	8.1	118.8	179.8	59.6	30	33.9			3.96	1.89	2.28		
LEU 30	8.52	120.8	177.6	58.4	40.8	27	25.8/22.5		3.96	1.89	1.46	0.84/0.72	
SER 31	7.98	113	177.2	61.6	62.7				3.96	3.86			
ALA 32	7.79	121.9	178.7	55.3	19.1				4	1.5			
THR 33	8.22	115.4	176.8	56.2	67.6	22.7			3.98	4.07	1.07		
VAL 34	8.32	121.8	178	66	31.5	22.4/20.4			3.49	1.69	0.84/0.60		
LYS 35	7.34	118.5	179.8	59.9	32.1	25.9	29.3	41.9	3.89	1.78	1.23	1.5	3.32
PHE 36	8.25	118.2	177.3	63.6	38.5		131.4	131.4	4.04	2.6		7.12	6.96
ARG 37	8.37	120.4	179.6	61.2	27.5	33.7	51.2						
GLN 38	8.65	119.3	179.3	59.6	27.5	34.2			3.46	1.63	2.76/2.32		
ILE 39	8.06	119.5	178	65.3	38	16.9/28.6	14		3.69	1.77	0.76/1.09	0.51	
LEU 40	7.73	116.8	177.1	55.3	42.4				4.2	1.85	1.58		
SER 41	7.67	115.1	173.5	58.4	63.7					3.87			
ARG 42	7.11	120.5	175.2	55.6	31.6	26.5	43.6		4.39	1.82	1.62	2.99	
ASP 43	8.29	119.7	176.6	55.5	41.5				4.35	2.63			
GLY 44	8.29	109.7	174.9	45.3					3.85				
GLU 46				59.7	29.1	36.4			4.08	2.08	2.32		
GLN 47	8.86	118.1	177	56.8	28	33.4			4.27	2.05	2.35		
ILE 48	7.28	119.6	177.6	66.3	38	18.8/29.6	17.3		3.34				
GLN 49	7.93	117.6	177.3	57.9	28	33.2			3.65	2.05	2.4		
ALA 50	7.5	119.8	180.5	55.1	17.8				4.08	1.39			
VAL 51	7.66	118.5	177.4	67.3	31.3	23.2/21.6			3.54	2.44	0.99/0.72		
ILE 52	7.8	119.8	181.5	65.5	37.9	16.5/28.8	13.9		3.61	1.96	0.88/0.76	0.75	
ASP 53	9.25	123	177.2	57.1	39.8								
ALA 54	7.55	119.9	178	52.5	18.6				4.24	1.5			
GLY 55	8.01	104.9	175.7	45.6					4.12/3.96				
ALA 56	6.8	117.8	177.9	53.8	19.8				3.88	0.88			
LEU 57	8.7	116.3	175.4	59.8	37.9								
PRO 58			179	66.3	30.6	28.3			4.12	2.32	1.81		
ALA 59	6.9	119.2	179.7	54.5	19.3				4.08	1.31			
LEU 60	8.03	118.9	178.2	58.1	42.3	26	25						
VAL 61	8	115.6	180	66.7	31.1	24.1/21.1			3.3	1.97	0.80/0.68		
GLN 62	7.88	121.4	178.8	58.9	28	33.9							
LEU 63	7.55	119.4	177.8	56.6	40.5	27.5	22.2		4.04	1.97	1.35		
LEU 64	7.49	115.4	177.1	56.1	40	25.8	22.4		4.16	1.96	1.4	0.72/0.64	
SER 65	7.48	111.9	174.3	57.9	64				4.53	3.92			
SER 66	7.67	118.7	174.8	56.5	64.7								
PRO 67			176.6	63	31.8	26.5	50.5		4.59	2.20/2.01			
ASN 68	8.21	120.2	174.8	53.4	39.8				4.63	3.02/2.63			
GLU 69	8.96	125.8	178	59.6	29.5	37.2			3.88	2.15	2.4		
GLN 70	8.1	119.9	178.1	59.4	28	34.6			4.05	2.16	2.4		
ILE 71	7.59	119.5	177.2	65.8	37.7	18.3/32.1	13.5		3.38	2.05	0.95/0.72	1	
LEU 72	7.9	117.7	178.7	58.6	43.3	25.7	24.4		3.88	1.86	1.34		

Residue	H	N	C	CA	CB	CG*	CD*	CE*	HA*	HB*	HG*	HD*	HE*
GLN 73	7.7	115.3	177.5	60.4	29.3	35.2			4.43	2.1			
GLU 74	7.53	117.6	178.9	58.9	29.8	35.2			4.31	1.75	1.97		
ALA 75	8.77	121.7	179.9	55.3	17.8				3.77	1.34			
LEU 76	8.52	117.1	178.9	58.1	43.8	33.9	26.8		3.81	2.31	2.17		
TRP 77	8.09	122.5	179	59.1	29		123.1	120.6	3.85	3.18		7.05	7.31/9.44
ALA 78	8.38	120.4	178.8	56.3	17.8				3.81	1.23			
LEU 79	8.7	115.8	178.8	58.4	42.8	26.6	24.7		3.73	1.85	1.23		
SER 80	8.62	114.6	176.8	63.5	67.2								
ASN 81	8.24	118.8	179.6	55.6	36.4					2.29			
ILE 82	8.26	122.4	177.2	65.7	37.7	17.3/29.8	14.1		3.69		0.68		
ALA 83	8.21	119.8	179.6	54.6	18.1				3.77	1.45			
SER 84	8.28	113.2	174.7	60.7	63.2					3.88/3.73			
GLY 85	8.02	110	174.3	46.4					3.92				
GLY 86	7.22	105.4	174.2	44.5					4.55/3.92				
ASN 87	8.81	117.4	176.6	57	39				4.19	2.87			
GLU 88	9.15	121.7	179.2	59.6	28.8	36.7			4	1.97	2.32		
GLN 89	8.43	122.6	177.7	60.1	26.2	34.9			3.85	1.96/1.81	2.32		
THR 90	7.88	115.4	176.9	69	68.2	24.6					1.34		
GLN 91	8.3	120.8	177	58.3	27.8	33.1			3.61	2.13/1.97	2.32		
ALA 92	7.6	120.6	180.4	55.3	17.1				4.12	1.42			
VAL 93	7.34	119.1	177.5	67	31.6	23.2/21.9			3.34	2.44	0.99/0.73		
ILE 94	7.8	119.9	181.2	65.8	37.9	16.3/28.5	13.7		3.53	1.97	0.90/1.38	0.72	
ASP 95	9.32	123.3	177	57.1	39.7				4.35	2.67			
ALA 96	7.41	119.7	176.9	52.3	18.6				4.31	1.34			
GLY 97	8.02	104.2	176.2	45.6					4.20/3.92				
ALA 98	7.08	117.1	178.6	55.1	20.4				3.81	1.11			
LEU 99	9.27	113.6	175.6	60.7	38.9								
PRO 100			179.3	65.8	30.6	28			4.24	2.3	1.81		
ALA 101	6.48	118.3	179.8	54.5	19.3				4.05	1.35			
LEU 102	8.24	118.6	178.4	57.9	42.3	26.3			3.92	2	1.15	0.72/0.57	
VAL 103	8.16	115.8	179.9	67	31.1	23.2/21.1			3.34	1.97			
GLN 104	7.63	119.9	178.8	58.9	28	33.9							
LEU 105	7.52	119.3	177.5	56.6	41.5	27.8	22.7		4.08	2.01	1.35		
LEU 106	7.36	115.7	176.8	56.1	40.2	25.7	21.9		4.2	1.93	1.42	0.72/0.60	
SER 107	7.36	111.8	174.3	57.9	63.8				4.51	3.92			
SER 108	7.58	118.9	174.6	56.9	65								
PRO 109			176.6	62.9	31.7	26.5	50.5		4.59	2.20/2.01		4.35	
ASN 110	8.21	120.5	175.2	52.8	39				4.59	2.95/2.59			
GLU 111	8.85	125.3	178.2	59.7	29.7	37			3.81	2.13	2.33		
GLN 112	8.13	120.1	177.6	59	28.3	34.4			3.92	1.81	1.98		
ILE 113	7.2	119.7	177.2	64.5	36.5				3.34		0.88		
LEU 114	8.08	118.3	178.5	58.4	43.3	27.2	25.2/24.3		3.85	1.85	1.24		
GLN 115	7.74	115.2	177.4	59.9	28	33.9			3.61	1.66	2.1		
TYR 116	7.99	117.1	177.6	59.4	37.2		130.7	117.8	4.24	3.10/2.91		6.99	6.53
ALA 117	8.98	123.7	179	55.5	17.3				4	1.35			
LEU 118	8.04	118.1		58.2	42.4	27.5	23.9/23.9		3.77	1.94	1.19		
ILE 119	7.94	119.5	178.8	65.3	38.3	16.9/27.9	14.1		3.73	1.62	0.88	0.48	
ALA 120	7.96	122.5	179.3	56.1	17.8				4.02	1.35			
LEU 121	8.22	117.2	179.6	58.1	42.5	26	24.9		3.65	1.89	1.39		
ASN 122	8.81	122.9	176.5	55.2	35.6								
ASN 123	8.54	120.3	178.3	57.2	39					2.87			
ILE 124	7.98	118.5	176.7	64.5	41				3.73	1.94			
ALA 125	8.11	119.8	179.7	54.9	19.5				3.73	1.58			
PHE 126	8.54	116.4	175.6	58.4	39.7		132.2	131.3	4.51	3.34/2.98		7.43	7.34
ALA 127	7.5	118.3	178.2	54.1	19				3.98	1.54			
GLY 128	7.42	103.1	173.9	45.6					3.87				
ASN 129	8.5	119.2	178	57.1	39				4.43	2.95/2.75			

Residue	H	N	C	CA	CB	CG*	CD*	CE*	HA*	HB*	HG*	HD*	HE*
GLU 130	9.21	121.1	179	59.6	28.5	36.7			3.96	1.89	2.25		
GLN 131	8.32	121.9	178	60.6	26.7	35.2			3.88				
THR 132	8.34	117	178.1	68.4	68.3	21.6							
GLN 133	8.26	120.9	177.3	58.3	27.5	33.1			3.73	2.16/2.05	2.36		
ALA 134	7.73	121.3	180.4	55.3	17.3				4.12	1.46			
VAL 135	7.49	120.1	177.5	67.5	31.3				4.01		1.04/0.76		
ILE 136	7.62	119.5	181.3	65.8	38	16.6/28.3	15		3.46	2.08			
ASP 137	9.18	122.9	177	57.1	39.8				4.35	2.67			
ALA 138	7.53	120.1	177	52.2	18.6				4.31	1.34			
GLY 139	8.03	104.2	176.1	45.6					4.19/3.96				
ALA 140	6.98	117	178.3	54.8	19.8				3.85	1.11			
LEU 141	9.19	114.9	175.2	61.2	39.3								
PRO 142			179.3	66	30.6	28.2			4.16	2.28	1.85		
ALA 143	6.72	119	180	54.3	19.3				4.04	1.35			
LEU 144	8.26	118.4	178.4	57.9	42.3	26	25						
VAL 145	8	115.6	179.1	67	31.1	23.7/21.4			3.3	2.01	0.84/0.72		
GLN 146	7.43	118	178.9	58.6	28	33.6							
LEU 147	7.55	119.8	177.7	56.6	41.3	27.3	23.4		4.08	2.17	1.5		
LEU 148	7.38	116.4	177	56.1	41.5	26.5	22.7		4.12	1.97	1.35	0.76/0.60	
SER 149	7.4	111.7	174.1	58.4	63.5				4.51	4			
SER 150	7.7	118.5	174.9	56.3	64.4								
PRO 151			176.8	63	31.8	26.5	50.7		4.67	2.24	2.01		
ASN 152	8.26	121.2	175.6	53.3	39.4								
GLY 153	9	114.4	175.6	47.4					3.61				
GLN 154	8.16	120.3	178.9	59.1	28	34.7			4.12	2.14/1.99	2.33		
ILE 155	7.57	120.6	179.7	64.7	36.7	17.3	13.2		3.41	2.13	0.84/1.50		
LEU 156	8.4	121.3	178.6	58.4	42.3	28.5	25.7/23.7		3.85	1.85	1.07	0.94/0.76	
GLN 157	8	116.9	179.3	60.4	29	34.9			3.83	2.09	2.39		
GLU 158	7.7	117.2	179.9	58.1	30.8	33.6			4.36	2.05			
THR 159	8.29	118.8	176.5	67.8	68.3	23.4			5.49	4.04	1.11		
LEU 160	8.69	121.8	179.7	58.6	41.3	27	25.7		3.77	1.93	1.58	1.22/0.68	
TRP 161	7.86	121.7	178.3	63.2	28.5		126.4		3.77	3.38/3.22		7.21	9.77
ALA 162	7.99	121.5	181.5	55.6	18.8				3.96	1.79			
LEU 163	8.45	115.8	178.3	57.6	42.8	27.3	24.2		3.81	1.74	1.15		
THR 164	7.94	115.1	175.3	67.1	67.8	21.6			3.92	3.95	0.99		
ASN 165	7.82	119.4	177.1	57.3	38.5				3.81	2.09			
ILE 166	7.34	117.6	176.9	65.8	38.5	17.3/30.0	13.5		3.3	1.62	1.77	0.64	
ALA 167	7.85	118.7	178.1	54.3	18.8				3.61	1.27			
MET 168	7.45	112	177.5	55.8	32.1	32.3		16.5	4.19	2.11/2.05	2.67/2.52		1.95
GLU 169	7.7	119.4	176.8	57.6	30.1	35.7			4.04	1.97	2.36/2.24		
GLY 170	7.06	103.1	173.4	45.1									
ASN 171			178	56.6	38.5				4.39	2.79			
GLU 172	9.12	121.1	179.1	59.9	28.5	36.7			3.96	1.89	2.25		
GLN 173	8.53	121.9	177.9	60.3	26.9	34.7							
LYS 174	8.13	119.6	178.8	60.7	32.4	26	29.8		3.65	2.16	1.7	1.89	
GLN 175	8.05	117.6	177.9	58.6	27.3	32.9			3.92	2.16/2.09	2.4		
ALA 176	7.78	122.4	180.9	55.1	17.3				4.07	1.4			
VAL 177	7.92	119.1	177.7	67	31.3	23.4			3.41	2.28	1.00/0.72		
LYS 178	7.78	118.9	181.4	60.2	31.9	26.2	29.5		3.88	1.97	1.22	1.7	2.83
GLU 179	8.95	121.8	177.4	58.9	29.3	36.7			4	2.05	2.44/2.28		
ALA 180	7.43	118.6	177.2	52.5	18.6				4.31	1.4			
GLY 181	7.75	103.6	176.3	45.4					4.28/3.92				
ALA 182	7.1	117.3	178.6	55.3	20.4				3.73	1.11			
LEU 183	8.81	115.7	178.8	59.6	40	27.5	25.2/23.9		3.69	1.7	1.5	0.87/0.72	
GLU 184	8.22	117.1	179.3	59.6	29	36.4			4	2.01/1.89	2.24		
LYS 185	6.75	115.8	179.6	57.7	33.1	24.7			4.08	1.78	1.46	1.7	2.65
LEU 186	8.34	120.3	179.3	58	41.8	26	23.9						

Residue	H	N	C	CA	CB	CG*	CD*	CE*	HA*	HB*	HG*	HD*	HE*
GLU 187	8.19	118.4	180.1	59.4	29.3	36.7			3.92	2.09	2.36/1.97		
GLN 188	7.28	117.4	178.6	58.1	28	33.7			4.08	2.17	2.52		
LEU 189	7.6	119.4	177.5	56.3	42.6	26.7	23.7		4.28	1.89	1.42		
GLN 190	7.42	115.8	175.8	58.5	28.8	34.9			4.08	2.17	2.52/2.32		
SER 191	7.48	110.9	174	57.1	63.5					3.88			
HIS 192	7.52	124.7	174.7	59.6	31.8				3.92	3.22/2.75			
GLU 193	7.87	123.6	176.8	58.5	29.4	35.7			3.96	1.89	2.12		
ASN 194	11.26	124.7	175.8	53.2	39.5				4.71	3.27/2.91			
GLU 195	9	127.6	177.9	59.7	29.5	36.2			3.92	2.05	2.3		
LYS 196	7.98	119.7	178.8	59.1	32.1	25	29	42.3	4.04	1.81	1.41/1.34	1.62	2.94
ILE 197	7.22	119.3	177.3	63.3	35.8	17.4/28.3	10.5		3.43	1.8	0.42/0.96	0.49	
GLN 198	8.08	118.1	178.9	59.6	28.5	33.6			3.92	2.16	2.60/2.32		
LYS 199	7.69	117.8	179.3	59.2	32.4		24.9						
GLU 200	8.12	119.6	180.1	59.1	29.4	36.4			4.04	1.93	2.44/2.32		
ALA 201	8.95	123.1	179	55.3	18.3				3.94	1.34			
GLN 202	8.08	118.6	178.8	59.4	28	33.6			4	2.17	2.52/2.36		
GLU 203	8.18	119.4	178.9	59.3	29.3	36.4			3.97	1.97	2.40/2.21		
ALA 204	7.92	121.5	179.2	55.3	18.1				3.88	1.27			
LEU 205	8.09	118.1	179	58.4	41.8	27	24.8		3.72	1.66	1.51		
GLU 206	7.98	118.5	179.9	59.4	29.1	36.4			3.92	2.17/2.05	2.36		
LYS 207	7.82	118	178.7	58.4	32.1	25.5	29		4	1.85	1.51	1.74	2.88
LEU 208	7.87	117.5	177.1	56.6	42.6	27	25.5/24.2		3.96	1.74	1.35	0.68/0.57	
GLN 209	7.6	115.7	176	56.1	29.6	34.1			4.28	2.01	2.4		
SER 210	7.67	116.1	173.5	58.4	64.1				4.31	3.85			
HIS 211	8.01	126.4	179.8	57.6	30.9								

Table 5.7 List of chemical shifts for backbone assignments of ^{2}H, ^{13}C, ^{15}N-labeled $Y_IMR^1R^2A_{II}$ in the presence of 2 equivalents NT.

Residue	H	N	C	CA	CB		Residue	H	N	C	CA	CB
GLN 16	7.54	117.6	178.6	58.4	27.3		SER 65	7.47	111.7	174.3	57.3	63
MET 17	8.13	119.9	178.3	58.4	33.5		SER 66	7.65	118.6	174.8	56.5	64.7
THR 18	8.38	115.3					ASN 68	8.22	120	174.8	52.7	39
GLN 19	7.5	119.9	179.5	58.3	27.2		GLU 69	8.95	125.5	178	59.1	28.6
GLN 20	7.88	120.1	177.9	58.4	28.2		GLN 70	8.08	119.6	178.1	59	27.2
LEU 21	7.49	116.5	177.2	56.8	41.2		ILE 71	7.59	119.5	177.2	65.2	36.6
ASN 22	7.32	113.3	174.7	52.1	39.3		LEU 72	7.9	117.4	178.7	58.1	42.3
SER 23	7.22	114.3	174.4	57.9	63		GLN 73	7.69	115	177.5	59.8	28.3
ASP 24	8.31	122.7	175.6	54.6	40.1		GLU 74	7.53	117.4	178.9	58.3	28.6
ASP 25	8.47	121.2	176.5	52.7	41.4		ALA 75	8.75	121.4	179.9	54.8	16.9
MET 26	8.87	126.8	178.2	58.7	32.4		LEU 76	8.5	116.9	178.9	57.6	
GLN 27	8.32	117.6	179.5	58.4	26.9		TRP 77	8.08	122.2	179	58.7	
GLU 28	7.66	121.8	178.9	58.7	28.6		ALA 78	8.37	120.1	178.8	55.7	16.8
GLN 29	8.1	118.5	179.8	59	29.1		LEU 79	8.67	115.6	178.8	57.7	41.5
LEU 30	8.5	120.6	177.6	57.9	39.8		SER 80	8.59	114.3		64.4	68
SER 31	7.98	112.8	177.2	61.1	61.9		ASN 81	8.22	118.5	179.6		
ALA 32	7.79	121.8	178.7	54.8	18.2		ILE 82	8.25	122.1	177.2	65.2	36.5
THR 33	8.21	115.1					ALA 83	8.19	119.5	179.6	54	17.1
VAL 34	8.29	121.5	178	65.3	30.5		SER 84	8.29	113.1	174.7	60.2	62.4
LYS 35	7.34	118.3	179.8	59.4	31.1		GLY 85	8.01	109.8	174.3	45.8	
PHE 36	8.24	117.9	177.3	63.1	37.9		GLY 86	7.22	105.2	174.2	44.2	
ARG 37	8.35	120.1	179.6	60.6			ASN 87	8.8	117.2	176.6	56.5	38.2
GLN 38	8.62	119.1	179.3	59	27		GLU 88	9.13	121.4	179.2	59.2	28.1
ILE 39	8.06	119.2	178	65	36.8		GLN 89	8.41	122.4	177.7	59.5	25.3
LEU 40	7.73	116.7	177.1	55.1	41.2		THR 90	7.86	115.2	176.9	67.4	68.3
SER 41	7.66	114.9					GLN 91	8.29	120.5	177	57.7	27
ARG 42	7.12	120.3	175.2	55.4	30.4		ALA 92	7.6	120.3	180.4	54.6	16.3
ASP 43	8.27	119.5	176.6	54.9	40.6		VAL 93	7.34	118.9	177.5	66.6	30.8
GLY 44	8.27	109.3	174.9	44.7			ILE 94	7.79	119.7	181.2	65.2	36.8
GLN 47	8.79	117.9					ASP 95	9.31	123.1	177	56.5	39
ILE 48	7.28	119.3	177.6	65.5	37.1		ALA 96	7.4	119.4	176.9	51.7	17.7
GLN 49	7.93	117.4	177.3	57.3	27		GLY 97	8.01	104.1	176.2	45	
ALA 50	7.52	119.6	180.5	54.6	16.8		ALA 98	7.1	116.9	178.6	54.6	19.6
VAL 51	7.66	118.3	177.4	66.6	30.5		LEU 99	9.26	113.3	175.6	60.1	37.6
ILE 52	7.8	119.6	181.5	65.2	36.9		ALA 101	6.47	118	179.8	53.8	18.5
ASP 53	9.24	122.7	177.2	56.5	39		LEU 102	8.23	118.3	178.4	57.3	41.2
ALA 54	7.55	119.7	178	52.1	17.7		VAL 103	8.16	115.6	179.9	66.3	30.2
GLY 55	8	104.7	175.7	45.1			GLN 104	7.63	119.7	178.8	58.3	27.2
ALA 56	6.82	117.6	177.9	53.2	18.8		LEU 105	7.51	119.1	177.5	56.2	40.4
LEU 57	8.69	116	175.4	59.8	37.9		LEU 106	7.35	115.5	176.8	55.4	39
ALA 59	6.91	119	179.7	54	18.5		SER 107	7.36	111.6	174.3	57.6	63.1
LEU 60	8.01	118.6	178.2	57.6	41.2		SER 108	7.57	118.8	174.6	56.5	64.1
VAL 61	8	115.4	180	66.1	30		ASN 110	8.21	120.3	175.2	52.4	38.2
GLN 62	7.87	121.2	178.8	58.1	27		GLU 111	8.84	125	178.2	59	28.9
LEU 63	7.54	119.2	177.8	56.2	39.5		GLN 112	8.12	119.9	177.6	58.4	27.5
LEU 64	7.48	115.2	177.1	55.7	39		ILE 113	7.2	119.6	177.2	63.9	35.4

Residue	H	N	C	CA	CB		Residue	H	N	C	CA	CB
LEU 114	8.07	118	178.5	57.8	42.2		LEU 163	8.44	115.6	178.3	57	41.7
GLN 115	7.72	114.9	177.4	59.5	27.2		THR 164	7.93	114.7	175.3	66.3	67.2
TYR 116	7.97	116.7	177.6		36.4		ASN 165	7.82	119.2	177.1	56.8	37.9
ALA 117	8.97	123.4	179	55.1	16.5		ILE 166	7.33	117.4	176.9	65.2	37.4
LEU 118	8.07	117.9					ALA 167	7.82	118.4	178.1	53.8	17.9
ILE 119	7.96	119.2					MET 168	7.45	111.9	177.5	55.4	
ALA 120	7.96	122.1	179.3	55.7	16.9		GLU 169	7.69	119.1	176.8	57	29.2
LEU 121	8.19	116.9					GLY 170	7.05	102.9	173.4	44.5	
ASN 122	8.8	122.5	176.5				GLU 172	9.11	120.8	179.1	59.2	27.8
ASN 123	8.52	119.9	178.3	56.5	38.4		GLN 173	8.51	121.6	177.9	59.5	26.1
ILE 124	7.96	118.3	176.7	63.9	37		LYS 174	8.12	119.4	178.8	60	31.3
ALA 125	8.08	119.5	179.7	54.3	18.5		GLN 175	8.04	117.4	177.9	57.9	26.4
PHE 126	8.51	116.1	175.6	57.9	39		ALA 176	7.78	122.1	180.9	54.6	16.6
ALA 127	7.5	118.1	178.2	53.5	18.2		VAL 177	7.91	118.9	177.7	66.3	30.5
GLY 128	7.41	102.9	173.9	45			LYS 178	7.77	118.7	181.4	59.8	30.8
ASN 129	8.49	119.1	178	56.8	38.2		GLU 179	8.94	121.7	177.4	58.4	28.3
GLU 130	9.19	120.9	179	58.9	27.8		ALA 180	7.43	118.4	177.2	51.9	17.7
GLN 131	8.29	121.6	178	59.8	25.9		GLY 181	7.74	103.4	176.3	44.7	
THR 132	8.33	116.8	178.1				ALA 182	7.11	117.2	178.6	54.6	19.3
GLN 133	8.24	120.6	177.3	57.6	26.4		LEU 183	8.79	115.4	178.8	59.2	39
ALA 134	7.72	121	180.4	54.6	16.3		GLU 184	8.19	116.9	179.3	59	28.1
VAL 135	7.49	119.9	177.5	66.9	30.5		LYS 185	6.76	115.7	179.6	57	32.2
ILE 136	7.62	119.3	181.3	65.2	37		LEU 186	8.32	120	179.3	57.5	40.7
ASP 137	9.17	122.7	177	56.5	39.2		GLU 187	8.18	118.1	180.1	59	28.3
ALA 138	7.53	119.9	177	51.7	17.9		GLN 188	7.28	117.2	178.6	57.6	27.5
GLY 139	8.01	104	176.1	45			LEU 189	7.6	119.2	177.5	56	41.4
ALA 140	6.99	116.8	178.3	54.3	19		GLN 190	7.42	115.6	175.8	57.9	28.1
LEU 141	9.17	114.7	175.2	60.6	38.2		SER 191	7.47	110.8	174	56.8	62.8
ALA 143	6.71	118.8	180	53.8	18.2		HIS 192	7.51	124.5	174.7	59.1	31.1
LEU 144	8.25	118.1	178.4	57	41.2		GLU 193	7.85	123.3	176.8	57.9	28.6
VAL 145	8	115.3	179.1	66.4	30		ASN 194	11.26	124.5	175.8	53	38.7
GLN 146	7.43	117.8	178.9	58.1	27.2		GLU 195	8.98	127.3	177.9	59	28.6
LEU 147	7.54	119.6	177.7	56.2	40.4		LYS 196	7.96	119.5	178.8	58.7	31.1
LEU 148	7.36	116.2	177	55.7	40.4		ILE 197	7.22	119.2	177.3	62.8	34.9
SER 149	7.4	111.6	174.1	57.9	62.9		GLN 198	8.06	117.8	178.9	59.2	27.6
SER 150	7.69	118.3	174.9	56.3	64.4		LYS 199	7.68	117.5	179.3	58.4	31.3
ASN 152	8.26	121	175.6	52.7	38.7		GLU 200	8.12	119.4	180.1	58.5	28.6
GLY 153	9	114.4	175.6	46.8			ALA 201	8.93	122.8	179	54.8	17.7
GLN 154	8.15	120.1	178.9	58.4	27.2		GLN 202	8.06	118.3	178.8	58.7	27.2
ILE 155	7.57	120.4	179.7	64.2	36		GLU 203	8.16	119.2	178.9	58.7	28.3
LEU 156	8.39	121	178.6	57.9	41.2		ALA 204	7.92	121.2	179.2	54.6	17.1
GLN 157	7.99	116.7	179.3	59.8	28.3		LEU 205	8.07	117.8	179	57.8	40.7
GLU 158	7.68	117	179.9	57.6	29.7		GLU 206	7.95	118.3	179.9	58.7	28.3
THR 159	8.28	118.6	176.5	62.2	67.2		LYS 207	7.8	117.8	178.7	57.9	31.1
LEU 160	8.67	121.5	179.7	58.1	40.1		LEU 208	7.86	117.2	177.1	55.9	41.3
TRP 161	7.85	121.5	178.3	62.6	28		GLN 209	7.59	115.5	176	55.4	28.6
ALA 162	7.99	121.3	181.5	55.1	17.9		SER 210	7.67	116.1	173.5	58.2	63.3
							HIS 211	8.01	126.2	179.8	57.1	30.2

Table 5.8 List of chemical shifts for backbone assignments of ^{2}H, ^{13}C, ^{15}N-labeled $Y_IMR^1R^2A_{II}$.

Residue	H	N	C	CA	CB		Residue	H	N	C	CA	CB
GLN 19	8.1	119.9	177.2	57	29		GLN 89	8.27	121.7	177.8		
GLN 20	7.91	120.2	176.9	56.2	31.7		THR 90	7.99	114.9	176.7		
LEU 21	7.88	121.6	177.5	55	41.1		GLN 91	8.13	120.4	177.1	57.8	27
ASP 24	8.31	122.4	175.7	54.5	40		ALA 92	7.63	120.5	180.5	54.7	16.5
ASP 25	8.44	121.1	176.6	52.9			VAL 93	7.41	118.4	177.4	66.5	30.7
GLN 27	8.32	117.6	179.5		26.9		ILE 94	7.7	120.1	181.2	65.2	36.9
GLU 28	7.64	121.4	178.8	58.6	28.7		ASP 95	9.25	123.2	177	56.7	39.1
GLN 29	8.05	118.8	177.3	58	27.3		ALA 96	7.38	119.3	176.8	51.8	17.9
LEU 30	8.37	119.7	178.2		40		GLY 97	7.97	104.1	176.2	45.1	
SER 31	7.97	113.2	175.4	61.1	62.1		ALA 98	7.13	117	178.5	54.5	19.4
ASP 43	8.18	119.1	176.8	54.5	40.7		LEU 99	9.27	113.1	175.6	59.9	37.7
GLY 44	8.13	108.6	174.6	45.2			ALA 101	6.48	118	179.8	54.1	18.4
ALA 50	7.55	120.1	178	53.5	18.3		LEU 102	8.22	118.4	178.5	57.5	41.2
VAL 51	7.33	118.5	177.3	66.6	30.7		VAL 103	8.18	115.9	179.8	66.4	30.2
ILE 52	7.83	120.4		61.5	37.3		GLN 104	7.63	119.7	178.9	58.3	27.3
ASP 53	9.14	122.9	177	56.7	39.2		LEU 105	7.49	119	177.6	56.3	40.4
GLY 55	8.02	105	175.7	45.2			LEU 106	7.37	115.4	176.8	55.6	39.2
ALA 56	6.9	117.7	178.1	50.7	18.8		SER 107	7.34	111.4	174.2	57.7	63
ALA 59	6.83	118.8	179.6	54.1	18.2		SER 108	7.54	118.6	174.6	56.4	64.2
LEU 60	7.95	118.7	178.4	57.5	41.2		ASN 110	8.2	120.1	175.2	52.3	38.2
VAL 61	8	115.7	179.9	66.1	30.1		GLU 111	8.83	125.4	178.1	59.2	29.1
GLN 62	7.82	120.7	178.8	58.2	27.2		GLN 112	8.1	118.9	177.8	58.6	27.5
LEU 63	7.52	119.2	177.8	56.3	39.6		ILE 113	7.19	119.7	177.3	63.7	35.6
LEU 64	7.44	115.5	177	55.7	39		LEU 114	8.05	118.5	178.7	57.9	42.1
SER 65	7.42	111.3	174.3	57.5	63		GLN 115	7.75	115.3	177.9	59.3	27.5
SER 66	7.57	118.2	174.8	56.4	64.5		TYR 116	7.78	115.8	178.5	60.1	37.3
ASN 68	8.21	120.1	174.7	52.6	39.2		ALA 117	8.95	123.8	178.9	55.1	16.9
GLU 69	8.93	125.8	177.8	59.3	28.8		LEU 118	8.19	118.5	178.4	57.5	41.4
GLN 70	8.04	119.3	178.2	58.9	27.3		ILE 119	7.94	118.8	178.7	65.1	37.1
ILE 71	7.51	120	177.8				ALA 120	7.86	120.8	179	55.4	16.9
LEU 72	7.93	118.7	178.9	58.3	42.2		LEU 121	8.37	117.5	179	57.8	41.2
GLN 73	7.82	115.2	177.8	59.4	28		ASN 122	8.66	119.4	177.4		
GLU 74	7.56	117.9	179.3	58.5	28.5		ILE 124	7.83	118.4	176.5	63.9	37.4
ALA 75	8.64	122.5	179.6	54.9	17.3		ALA 125	8.21	118.6			18
LEU 76	8.42	117.7	179.3	57.5	41.5		GLY 128	7.47	103.9	173.9	44.9	
TRP 77	8.28	121.9	179.3	60.4			ASN 129	8.48	119.1	177.4	57.8	38.2
ALA 78	8.38	121.5	179.3	55.3	16.8		GLU 130	9.06	120.7	178.7	59.1	27.7
LEU 79	8.5	116.7	178.8	57.7	41.2		GLN 131	8.37	121	177.8	52.9	26.1
SER 80	8.27	113.6	175		61.9		GLN 133	8.29	120.6	177.3	57.8	26.7
ILE 82	8.11	120.7	177.6	64.5	37.1		ALA 134	7.71	120.9	180.5	54.7	16.7
ALA 83	8.06	118.3		53.8	17.3		VAL 135	7.55	119.2	177.4	66.7	30.7
SER 84	7.48	111.2		59.4	62.7		ILE 136	7.74	120	181.2	65.3	36.9
GLY 85	7.86	108.7	174	45.7			ASP 137	9.14	122.9	177	56.7	39.2
GLY 86	7.34	105.7	174.3	44.3			ALA 138	7.46	119.5	176.9	51.8	17.9
GLU 88	9.14	120.8	179.1	59.3	27.9		GLY 139	7.97	104	176.2	45.1	

Residue	H	N	C	CA	CB		Residue	H	N	C	CA	CB
ALA 140	7.1	116.9	178.5	54.5	19.2		LYS 178	7.85	118.9	181.4	59.7	31
LEU 141	9.21	113.9	175.5	60.3	37.8		GLU 179	8.91	121.6	177.3	58.4	28.3
ALA 143	6.58	118.5	180	54	18.3		ALA 180	7.41	118.5	177.1	51.9	17.8
LEU 144	8.27	118.2	178.5	57.1	41.1		GLY 181	7.75	103.4	176.3	44.8	
VAL 145	8.12	115.9	179.2	66.6	30.2		ALA 182	7.09	117.2	178.6	54.7	19.4
GLN 146	7.46	117.9	178.9	58.1	27.2		LEU 183	8.75	115.3	178.8	59.1	39.1
LEU 147	7.5	119.7	177.8	56.3	40.4		GLU 184	8.12	117	179.3	59.1	28.1
LEU 148	7.37	116.2	176.9	55.7	40.1		LYS 185	6.77	115.7	179.4	57	32
SER 149	7.36	111.3	174.2	57.9	63		LEU 186	8.3	119.9	179.2	57.7	40.8
SER 150	7.67	118.2	174.9	56.5	64.4		GLU 187	8.14	117.9	180.1	59	28.4
ASN 152	8.28	121.1	175.5	52.7	38.8		GLN 188	7.27	117.3	178.6	57.8	27.4
GLY 153	8.91	114.1	175.4	47			LEU 189	7.66	119.3	177.5	55.9	41.4
GLN 154	8.08	121	178.7	58.6	27.2		GLN 190	7.39	115.6	175.8	58.1	28.1
ILE 155	7.55	120.7	179.6	64.1	36.1		SER 191	7.47	110.8	173.9	56.7	62.8
LEU 156	8.39	121.4	178.5	58	41.4		HIS 192	7.49	124.5	174.7	59.1	31.2
GLN 157	7.93	116.7	178.5	59.4	28.3		GLU 193	7.85	123.3	176.9	58.1	28.6
GLU 158	7.72	117.1	179.6	58.1	29.1		ASN 194	11.27	124.5	175.9	53	38.9
THR 159	8.28	119.1	176.5				GLU 195	8.97	127.4	177.9	59.1	28.7
LEU 160	8.52	122	179.7	58.2	40.4		LYS 196	7.96	119.4	178.9	58.9	31.1
TRP 161	7.95	121.1	178.3	54.8			ILE 197	7.2	119.2	177.3	62.7	34.8
ALA 162	8.08	122	180.7	55.2	17.7		GLN 198	8.06	117.9	178.9	59.2	27.7
LEU 163	8.4	115.3	178.5	57.4	41.6		LYS 199	7.65	117.5	179.3	58.6	31.4
THR 164	8.01	113.8		66.5	67.5		GLU 200	8.07	119.3	180	58.6	28.4
ASN 165	7.85	119	177.5	56.4	37.7		ALA 201	8.91	122.6	179	54.8	17.6
ILE 166	7.7	118.5	176.9	65.3	37.4		GLN 202	8.05	118.2	178.7	58.9	27.4
ALA 167	7.75	118.1	177.9	53.8	18.1		GLU 203	8.12	119.2	178.9	58.8	28.5
MET 168	7.46	111.7	177.3	55.5	31.5		ALA 204	7.96	121.3	179.2	54.8	17.3
GLU 169	7.62	118.9	176.7	57.2	29.2		LEU 205	8.09	117.9	179	57.9	40.8
GLY 170	7.06	103	173.4	44.7			GLU 206	7.92	118.2	179.9	58.9	28.4
GLU 172	9.13	120.7	179.2	59.3	27.7		LYS 207	7.78	117.8	178.7	58.1	31.2
GLN 173	8.57	121.8	177.8	59.8	26.2		LEU 208	7.89	117.3	177.2	56.2	41.4
LYS 174	8.16	119.7	178.7	60.1	31.1		GLN 209	7.62	115.5	176.1	55.6	28.7
GLN 175	8.07	117.5	177.9	58.1	26.6		SER 210	7.64	116	173.5	58.2	63.4
ALA 176	7.78	122	180.9	54.7	16.7		HIS 211	8	126.1	179.7	57.2	30.2
VAL 177	7.93	119.1	177.7	66.7	30.6							

Table 5.9 List of chemical shifts for backbone assignments of ^{2}H, ^{13}C, ^{15}N-labeled $Y_IMR^1R^2\ R^3MA_{II}$ in the presence of 2 equivalents of NT.

Residue	H	N	C	CA	CB		Residue	H	N	C	CA	CB
GLN 20	7.88	120	177.8	58.4	28.2		LEU 202	8.72	118.5	179.2	57.5	40.9
LEU 21	7.48	116.3	177.2	57	41.2		ASP 203	8.38	120.9	178.3	56.8	40
ASN 22	7.31	113.1	174.6	52.1	39.3		ALA 204	7.76	120.5	179.7	54.6	17.1
SER 23	7.22	114.2	174.3	58.1	63		LEU 205	8.2	116.5	178.6	57.8	41.4
ASP 24	8.31	122.6	175.6	54.5	40.1		THR 206	8.49	113.6	178.6	66.5	67.6
ASP 25	8.47	121.1	176.4	52.6	41.4		ASN 207	7.7	118.8	177.5	56	38.1
MET 26	8.86	126.7	178.1	58.8	32.4		ILE 208	7.64	118.9	176.7	64.1	36.9
GLN 27	8.31	117.4	179.4	58.6	26.9		ALA 209	7.89	119.1	177.4	53.3	16.9
GLU 28	7.64	121.6	178.8	58.6	28.6		GLU 210	7.34	114.2	177.4	57	29.2
GLN 29	8.08	118.4	179.7	58.9	29		LEU 211	7.6	118.1	177.5	56.2	40.8
LEU 30	8.49	120.4	177.6	57.9	39.8		GLY 212	7.16	103.9	173.6	44.3	
SER 31	7.96	112.7		61.2	61.9		GLY 223	7.96	103.8	176.2	45.1	
GLU 69	8.93	125.4	177.9	59.2	28.7		ALA 224	7.28	116.9	178.6	54.5	19.6
GLY 85	8	109.8	174.3	45.9			LEU 225	9.28	112.8	175.6	59.8	37.4
GLY 86	7.21	105	174.2	44.2			ALA 227	6.52	118	179.8	54	18.5
ASN 110	8.2	120.2	175.1	52.5	38.3		LEU 228	8.26	118.1	178.4	57.3	41.3
GLU 111	8.82	124.9	178.1	59.2	28.9		VAL 229	8.23	116	179.4	66.3	30.3
GLN 112	8.1	119.7	177.5	58.5	27.5		LEU 232	7.39	116	176.5	55.7	39.2
ILE 113	7.18	119.4	177.1	63.8	35.6		SER 233	7.31	110.4	174	57.4	63
LEU 114	8.06	117.8	178.4	57.8	42.2		SER 234	7.46	117.8	174.6	56.4	64.1
GLN 115	7.71	114.8	177.3	59.5	27.2		ASN 236	8.22	119.9	174.9	52.3	38
GLY 128	7.39	102.7	173.8	45			GLU 237	8.84	125.5	177.9	59.2	29
ILE 155	7.51	120	179.6	64.1	36.1		GLN 238	8.07	119.3		58.8	27.2
LEU 156	8.47	121.2	178.5	57.8	40.1		ILE 239	7.2	121	176.9	63.3	35.9
GLN 157	8.04	116.9	178.3	59.9	27.6		LEU 240	8.06	120	178.9	58.7	42.1
GLU 158	7.54	115.7	180	57.7	29.9		GLN 241	7.78	115	177.5	59	28.1
ALA 167	7.85	119	177.9	54	17		GLU 242	7.65	118.6	179	58.7	28.1
MET 168	7.49	112.3	177.5	55.9	31.7		ALA 243	8.64	123	178.9	54.9	17.3
GLU 169	7.71	118.3	177	56.9	29.4		LEU 244	8.26	118.9	178.7	58	40.8
GLY 170	7.24	103.7	173.5	44.5			TRP 245	8.3	120.7	178.8	59	26.8
ALA 185	6.75	118.5	180.1	54.1	18.6		ALA 246	8.22	121.5	180	55.1	16.9
LEU 186	8.18	118.8	178.6	57.7	41		LEU 247	8.29	116.9	178.4	57.6	41.2
VAL 187	8.1	116.1	179.7	66.2	30		SER 248	8.61	114.1	177.2	61	62.5
LEU 189	7.43	118.3	177.7	56.2	40.2		ASN 249	7.95	118.5	178.4	55.9	37.4
LEU 190	7.45	115.4	176.7	55.7	39.3		ILE 250	8.01	121.6	177.3	64.7	36.7
SER 191	7.3	110.8	174.1	57.4	63		ALA 251	8.18	118.4	176.5	53.2	17.6
SER 192	7.48	118	174.7	56.4	64.2		SER 252	7.27	111	174.6	59.3	63.1
ASN 194	8.17	119.8	174.8	52.4	38.5		GLY 253	7.84	108.3	172.8	45	
GLU 195	8.89	125.9	177.6	59.3	28.9		GLY 254	7.52	104.4	174.3	44.1	
GLN 196	7.97	118.4	177.9	58.6	27.6		GLN 257	7.72	121.2	177.4	59.7	26.2
ILE 197	7.32	119.7	177.1	63.7	35.8		LYS 258	8.15	118.8	178.8	60.1	31.7
LEU 198	7.95	119.1	178.7	58.3	41.8		GLN 259	8.09	117.2	177.7	58.1	26.6
GLN 199	7.84	116.1	178.5	59.2	27.9		ALA 260	7.55	121.3	180.6	54.6	16.7
TYR 200	7.78	117.4	179.2	59.5	38.3		VAL 261	7.72	118.7	177.7	66.7	30.3
ALA 201	8.51	122	178.7	55	17.7		LYS 262	7.85	118.5	181.1	59.9	30.9

Residue	H	N	C	CA	CB
GLU 263	8.75	121	177.1	58.2	28.3
ALA 264	7.33	118.5	176.8	51.7	18
GLY 265	7.71	103.4	176.2	44.9	
ALA 266	7.01	116.9	178.5	54.6	19.3
LEU 267	8.8	115.5		59.2	39.4
GLU 268	8.12	116.4	179.3	59.1	28.2
LYS 269	6.73	115.4	179.1	56.6	31.5
LEU 270	8.34	119.2	179.1	57.7	40.6
GLU 271	8.16	117.2	180.1	58.9	28.5
GLN 272	7.3	117.9	178.8	57.9	27.4
LEU 273	7.8	119.3	177.5	56	41.3
GLN 274	7.41	115.7	175.7	58.1	28.1
SER 275	7.49	110.6	173.9	56.8	62.9
HIS 276	7.52	124.8	174.6	59.2	31.2
GLU 277	7.87	123.7	177.1	58.3	28.9
ASN 278	11.48	124.2	176.2	52.8	38.1
GLU 279	8.93	127.3	178	59	28.8
LYS 280	8.13	119.4	179.2	58.9	31
ILE 281	7.11	119.7	177	63.8	35.5
GLN 282	8	117.5	178.9	59.2	27.7
LYS 283	7.67	117.3	179.2	58.7	31.4
GLU 284	7.97	119.1	179.8	58.4	28.4
ALA 285	8.84	122.1	178.8	54.9	17.5
GLN 286	8.03	117.7	178.6	58.9	27.3
GLU 287	8.1	118.9	179.1	58.9	28.5
ALA 288	8.06	121.3	179.1	54.8	16.9
LEU 289	8.2	118.3	179	57.8	40.8
GLU 290	7.93	118.2	179.8	58.9	28.3
LYS 291	7.67	118	178.8	58.7	31.7
LEU 292	7.75	116.8	177	56	41.6
GLN 293	7.62	115.5	175.9	55.6	28.7
HIS 295	7.95	126.1	179.9	57.3	30.4

Table 5.10 List of chemical shifts for backbone assignments of ^{2}H, ^{13}C, ^{15}N-labeled $Y_IMR^1R^2\ R^3MA_{II}$

Residue	H	N	C	CA	CB*		Residue	H	N	C	CA	CB*
GLN 73	7.8	115.1	175.1	56.9	25.4		GLN 238	8.07	119.2	175.6	56.1	24.5
GLU 74	7.56	117.8	176.7	55.9	25.9		ILE 239	7.19	121	174.4	60.8	33.4
ALA 75	8.64	122.3	177	52.3	14.6		LEU 240	8.05	120	176.4	56.1	39.4
LEU 76	8.42	117.5	176.8	54.9	38.9		GLN 241	7.78	115.1	175	56.6	25.5
TRP 77	8.26	121.8	176.8	57.9			GLU 242	7.65	118.6	176.5	55.9	25.5
ALA 83	8.05	118.2		51.1	14.9		ALA 243	8.64	123	176.3	52.3	14.7
SER 84	7.49	111.1		56.8	60.1		LEU 244	8.26	119	176.1	55.5	38.2
GLY 85	7.86	108.6	171.4	43.2			TRP 245	8.3	120.7	176.3	56.5	24.1
GLY 86	7.34	105.6	171.7	41.7			ALA 246	8.22	121.5	177.5	52.5	14.4
ASN 110	8.21	120	172.5	49.7	35.5		LEU 247	8.3	116.9	175.9	55	38.6
GLU 111	8.83	125.3	175.4	56.6	26.4		SER 248	8.61	114.1	174.7	58.5	60
GLN 112	8.1	118.8	175.2	56.2	24.8		ASN 249	7.95	118.4	175.9	53.3	34.8
ILE 113	7.18	119.5	174.7	61.1	33		ILE 250	8.01	121.6	174.7	62.2	34.2
LEU 114	8.05	118.3		55.3	39.5		ALA 251	8.18	118.4	173.9	50.6	15.1
GLN 115	7.74	115.2	175.2	56.7	24.8		SER 252	7.27	111.1		56.8	60.5
TYR 116	7.78	115.5	175.9	57.4	34.7		GLY 253	7.84	108.3	170.3	42.4	
GLN 154	8.07	120.7	176.1	56	24.6		GLY 254	7.53	104.4	171.8	41.5	
ILE 155	7.52	120.4	177.1	61.5	33.5		GLN 257	7.73	121.2	174.9	57.1	23.5
LEU 156	8.47	121.3	175.9	55.2	38.9		LYS 258	8.15	118.8	176.2	57.4	29.2
GLN 157	8.04	116.9	175.4	57.2	25.1		GLN 259	8.09	117.2	175.2	55.5	24
GLU 158	7.62	116.4	177.1	55.4	26.8		ALA 260	7.55	121.3	178	52.1	14.2
ALA 167	7.81	118.6	175.3	51.4	14.6		VAL 261	7.71	118.4	175.1	64.1	27.7
MET 168	7.45	112.1	174.8	53.4	29.2		LYS 262	7.85	118.5	178.5	57.3	28.3
GLU 169	7.63	118.3	174.3	54.4	26.7		GLU 263	8.75	121	174.5	55.6	25.7
GLY 170	7.22	103.6	170.9	42			ALA 264	7.33	118.5	174.3	49.2	15.4
ASN 194	8.16	119.8	172.3	49.7	35.8		GLY 265	7.71	103.4	173.7	42.3	
GLU 195	8.86	125.8	175.1	56.6	26.3		ALA 266	7.01	116.9	176	52.1	16.8
GLN 196	7.98	118.5	175.2	55.7	25		LEU 267	8.81	115.5	176	56.6	36.8
ILE 197	7.26	119.6	174.6	60.9	33		GLU 268	8.13	116.4	176.7	56.5	25.7
LEU 198	7.95	119.3	176.2	55.7	39.3		LYS 269	6.73	115.4	176.6	54	28.9
GLN 199	7.82	116.1	175.9	56.5	25.3		LEU 270	8.34	119.2	176.5	55.1	38
TYR 200	7.79	117.5	176.7	56.7	35.4		GLU 271	8.15	117.2	177.6	56.4	25.9
ALA 201	8.56	122.3	176.2	52.5	15		GLN 272	7.3	117.9	176.3	55.3	24.8
LEU 202	8.69	118.6	176.7	54.9	38.2		LEU 273	7.8	119.3	175	53.3	38.7
ASP 203	8.4	120.9	175.7	54.3	37.5		GLN 274	7.4	115.7	173.1	55.5	25.5
ALA 204	7.77	120.6	177.2	52	14.6		SER 275	7.49	110.6	171.4	54.2	60.3
LEU 205	8.2	116.6	176.1	55.2	38.8		HIS 276	7.52	124.8	172.1	56.6	28.6
THR 206	8.5	113.6	174	64	65		GLU 277	7.87	123.7	174.6	55.7	26.3
ILE 208	7.64	119	176.6	61.5	34.4		ASN 278	11.48	124.1	173.6	50.2	35.5
ALA 209	7.89	119.1	174.8	50.6	14.3		GLU 279	8.93	127.3	175.4	56.4	26.1
GLU 210	7.35	114.3	174.8	54.4	26.6		LYS 280	8.13	119.4	176.6	56.3	28.4
LEU 211	7.61	118.2	174.9	53.6	38.2		ILE 281	7.11	119.7	174.5	61.2	33
GLY 212	7.19	104	171.1	41.8			GLN 282	8	117.5	176.4	56.5	25.2
ASN 236	8.22	119.9	172.4	49.7	35.4		LYS 283	7.67	117.3	176.7	56.1	28.8
GLU 237	8.85	125.5	175.4	56.6	26.4		GLU 284	7.97	119.2	177.3	55.8	25.8

Residue	H	N	C	CA	CB*
ALA 285	8.85	122.1	176.3	52.3	14.9
GLN 286	8.04	117.7	176.1	56.3	24.7
GLU 287	8.1	118.8	176.6	56.2	25.9
ALA 288	8.06	121.3	176.6	52.2	14.3
LEU 289	8.19	118.3	176.5	55.2	38.3
GLU 290	7.93	118.2	177.2	56.3	25.7
LYS 291	7.67	118.1	176.3	56.1	29.1
LEU 292	7.75	116.8	174.5	53.4	39.1
GLN 293	7.62	115.6	173.4	53	26
HIS 295	7.95	126.1	177.3	54.7	27.8

References

1. Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* **23**, 1105–1116.
2. Binz, H. K., Amstutz, P. & Plückthun, A. (2005). Engineering novel binding proteins from nonimmunoglobulin domains. *Nature biotechnology* **23**, 1257–1268.
3. Boersma, Y. L. & Plückthun, A. (2011). DARPins and other repeat protein scaffolds. advances in engineering and applications. *Current opinion in biotechnology* **22**, 849–857.
4. Caravella, J. & Lugovskoy, A. (2010). Design of next-generation protein therapeutics. *Current opinion in chemical biology* **14**, 520–528.
5. Hosse, R. J., Rothe, A. & Power, B. E. (2006). A new generation of protein display scaffolds for molecular recognition. *Protein science : a publication of the Protein Society* **15**, 14–27.
6. Löfblom, J., Frejd, F. Y. & Stahl, S. (2011). Non-immunoglobulin based protein scaffolds. *Current opinion in biotechnology* **22**, 843–848.
7. Almagro, J. C. (2004). Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size. implications for the rational design of antibody repertoires. *Journal of molecular recognition : JMR* **17**, 132–143.
8. Kuriyan, J. & Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Annual review of biophysics and biomolecular structure* **26**, 259–288.
9. Esteban, O. & Zhao, H. (2004). Directed evolution of soluble single-chain human class II MHC molecules. *J. Mol. Biol.* **340**, 81–95.
10. Coates, J. C. (2003). Armadillo repeat proteins. beyond the animal kingdom. *Trends in cell biology* **13**, 463–471.
11. Blatch, G. L. & Lassle, M. (1999). The tetratricopeptide repeat. a structural motif mediating protein-protein interactions. *BioEssays : news and reviews in molecular, cellular and developmental biology* **21**, 932–939.
12. Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. (1999). The WD repeat. a common architecture for diverse functions. *Trends in biochemical sciences* **24**, 181–185.
13. Andrade, M. A. & Bork, P. (1995). HEAT repeats in the Huntington's disease protein. *Nat. Genet.* **11**, 115–116.
14. Bennett, V. & Stenbuck, P. J. (1979). Identification and partial purification of ankyrin, the high affinity membrane attachment site for human erythrocyte spectrin. *J. Biol. Chem.* **254**, 2533–2541.
15. Hatzfeld, M. (1999). The armadillo family of structural proteins. *International review of cytology* **186**, 179–224.
16. Anastasiadis, P. Z. & Reynolds, A. B. (2000). The p120 catenin family: complex roles in adhesion, signaling and cancer. *J. Cell. Sci.* **113 (Pt 8)**, 1319–1334.
17. Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). Three-dimensional structure of the armadillo repeat region of beta-catenin. *Cell* **90**, 871–882.
18. Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. (1998). Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. *Cell* **94**, 193–204.
19. Riggleman, B., Wieschaus, E. & Schedl, P. (1989). Molecular analysis of the armadillo locus. uniformly distributed transcripts and a protein with novel internal repeats are associated with a Drosophila segment polarity gene. *Genes & development* **3**, 96–113.
20. Kippert, F. & Gerloff, D. L. (2009). Highly sensitive detection of individual HEAT and ARM repeats with HHpred and COACH. *PloS one* **4**, e7148.

21. Huber, A. H. & Weis, W. I. (2001). The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. *Cell* **105**, 391–402.
22. Catimel, B., Teh, T., Fontes, M. R., Jennings, I. G., Jans, D. A., Howlett, G. J. *et al. (*2001). Biophysical characterization of interactions involving importin-alpha during nuclear import. *The Journal of biological chemistry* **276**, 34189–34198.
23. Daniels, D. L. & Weis, W. I. (2002). ICAT inhibits beta-catenin binding to Tcf/Lef-family transcription factors and the general coactivator p300 using independent structural modules. *Molecular cell* **10**, 573–584.
24. Varadamsetty, G., Tremmel, D., Hansen, S., Parmeggiani, F. & Plückthun, A. (2012). Designed Armadillo repeat proteins. library generation, characterization and selection of peptide binders with high specificity. *Journal of molecular biology* **424**, 68–87.
25. Parmeggiani, F., Pellarin, R., Larsen, A. P., Varadamsetty, G., Stumpp, M. T., Zerbe, O. *et al. (*2008). Designed armadillo repeat proteins as general peptide-binding scaffolds. consensus design and computational optimization of the hydrophobic core. *Journal of molecular biology* **376**, 1282–1304.
26. Alfarano, P., Varadamsetty, G., Ewald, C., Parmeggiani, F., Pellarin, R., Zerbe, O. *et al. (*2012). Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy. *Protein science : a publication of the Protein Society* **21**, 1298–1314.
27. Nieto, J. L., Rico, M., Santoro, J., Herranz, J. & Bermejo, F. J. (1986). Assignment and conformation of neurotensin in aqueous solution by 1H NMR. *International journal of peptide and protein research* **28**, 315–323.
28. Hajduk, P. J., Meadows, R. P. & Fesik, S. W. (1999). NMR-based screening in drug discovery. *Q. Rev. Biophys.* **32**, 211–240.
29. Wetzel, S. K., Ewald, C., Settanni, G., Jurt, S., Plückthun, A. & Zerbe, O. (2010). Residue-resolved stability of full-consensus ankyrin repeat proteins probed by NMR. *Journal of molecular biology* **402**, 241–258.
30. Sambrook, J. & Russell, D. W. (2001). Molecular cloning. A laboratory manual, 3rd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
31. Aslanidis, C. & Jong, P. J. de (1990). Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res.* **18**, 6069–6074.
32. Watson, R. P., Bumbak, F., Ewald, C., Reichen, C., Plüchthun, A. & Zerbe, O. (2013). Spontaneous Self-Assembly of Fragments of Engineered Armadillo Repeat Proteins into Folded Proteins, unpublished.
33. van den Berg, S., Löfdahl, P.-A., Härd, T. & Berglund, H. (2006). Improved solubility of TEV protease by directed evolution. *J. Biotechnol.* **121**, 291–298.
34. Reichen, C., Hansen, S. & Plüchthun, A. (2013). Modular Peptide Binding: From a comparison of natural binders to designed Armadillo Repeat Proteins, accepted.
35. Cornilescu, G., Delaglio, F. & Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13**, 289–302.
36. FIELDING, L. (2007). NMR methods for the determination of protein–ligand dissociation constants. *Progress in Nuclear Magnetic Resonance Spectroscopy* **51**, 219–242.
37. Christen, M. T., Menon, L., Myshakina, N. S., Ahn, J., Parniak, M. A. & Ishima, R. (2012). Structural basis of the allosteric inhibitor interaction on the HIV-1 reverse transcriptase RNase H domain. *Chem Biol Drug Des* **80**, 706–716.
38. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293.

39. Webb, C., Upadhyay, A., Giuntini, F., Eggleston, I., Furutani-Seiki, M., Ishima, R. *et al. (*2011). Structural features and ligand binding properties of tandem WW domains from YAP and TAZ, nuclear effectors of the Hippo pathway. *Biochemistry* **50**, 3300–3309.
40. Bussi, G., Donadio, D. & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J Chem Phys* **126**, 14101.
41. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F. & DiNola, A. a. J. R. H. (1984). Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **81**, 3684–3691.
42. Darden, T., York, D. & Pedersen, L. (1993). Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *Journal of Chemical Physics* **98**, 10089–10092.
43. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. (1997). LINCS: A linear constraint solver for molecular simulations. *Journal of Chemical Physics* **18**, 1463–1472.
44. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4**, 435–447.
45. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118**, 11225–11236.
46. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **79**, 926–935.
47. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **3**, 171–176.
48. Conti, E. & Kuriyan, J. (2000). Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure* **8**, 329–338.
49. Madhurantakam, C., Varadamsetty, G., Grütter, M. G., Plückthun, A. & Mittl, P. R. (2012). Structure-based optimization of designed Armadillo-repeat proteins. *Protein science : a publication of the Protein Society* **21**, 1015–1028.
50. Lin, Y. & Wagner, G. (1999). Efficient side-chain and backbone assignment in large proteins: application to tGCN5. *J. Biomol. NMR* **15**, 227–239.